THE IMPACT OF XML IN DIGITAL LIBRARY DEVELOPMENT

Naicheng Chang

Thesis Submitted for the Degree of
Doctor of Philosophy of the University of London

School of Library, Archive and Information Studies
University College London
2005

UMI Number: U593039

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U593039
Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This thesis examines the strength of Extensible Markup Language (XML) technology and explores the impact of the implementation of XML in three digital library initiatives which are used as case studies by examining in depth their digital library development.

In part I, metadata issues are covered and the different kinds of metadata and metadata systems available for use in digital libraries are evaluated. Different models of storing the data and their relationship with XML are evaluated. The role of SGML in the early development of digital libraries is covered. To illustrate data manipulation, the author experiments with a digital library of images and their evaluations which use an XML-aware relational database. This part also concludes content management of both text and non-text which have used XML or may in the future move towards XML.

In part II, three case studies are examined and the results of interviews done by the author when visiting them assist in a detailed evaluation of the three examples. The three digital libraries are: the Library of Congress National Digital Library Program (NDLP), the University of Michigan Digital Library Services (DLS) and the Perseus Digital Library (PDL). Each digital library is evaluated in depth taking into account the digitization processes used, the metadata systems they employ and how the metadata are created, and the delivery systems. The HTTP usage statistics of the three case studies are also analyzed. The on-going work required for the maintenance of the digital libraries, managerial aspects relating to staff management, information on users and their usage and organizational aspects are covered and evaluated.

The author evaluates in what ways in each of those aspects the use of XML could benefit the digital library's development. The thesis finishes with a number of recommendations that could be taken up by digital libraries to their benefit.

To my husband and children

Acknowledgements

I would like to thank my supervisor Professor Susan Hockey for her encouragement and guidance, without which this research would never have started. Also, I wish to thank her for her help in connection with the Research Projects Fund of the Graduate School at UCL and her personal connection to the three case study digital libraries, without which this research would never have finished. I would like to thank Claire Warwick for her feedback and comments which sparkled different ways of thinking on the thesis; also, additional thanks to Claire for continuing to supervise me with no obligation after Professor Hockey had retired so that I could finalize the thesis. I would like to thank Melissa Terras for her comments. I would also like to thank Elizabeth Danbury for her comments and encouragement. I would like to thank Professor Ia McIlwaine and Professor John McIlwaine for their warm welcome when I just arrived in London. Very special thanks go to Professor Ia McIlwaine who provided her strong and warm assistance at the time when my emotion was at its lowest ebb.

Special thanks go to Kerstin Michaels who was very patient with my tedious enquiries. Also, many thanks go to the kind assistance and attention from Laura Keshav at the Departmental Office. In particular, I would like to thank Eccy de Jonge who provided me with more than I could expect as a research student at SLAIS.

This research interviews have been developed with the kind assistance of many people. I am indebted to research members of the Perseus team including Professor Gregory Crane, Clifford Wulfman, Anne Mahoney, Thomas Milbank and former team member David Smith; staff members in the University of Michigan Library, John Price-Wilkin, Christina Powell and John Weise; staff members in The Library of Congress, Sally McCallum, Caroline Arms, Martha Anderson, Carl Fleischhauer, Laura Graham, Steven McCollum, David Woodward and Mary Ambrioso. Without their help, the case studies could not possibly have been accomplished.

I would like to thank my parents, my brothers and sister, without their encouragement, this would never been possible. Most special thanks to my dearest husband and children for their love and support in allowing me to fulfil my dream. I would like to thank Alan Hopkinson's hospitality in inviting my family to spend a traditional English Boxing Day with his family.

The research case studies of this thesis were partially supported by the Graduate School Research Projects Fund at University College London. I am grateful for its support.

Contents

List of Figures

List of Abbreviations

AAP	Association of American Publishers
ACH/ALLC	Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing
ACM	Association for Computing Machinery
AHDS	Arts and Humanities Data Service
AHRB	Arts and Humanities Research Board
ALA	American Library Association
ALCTS	Association for Library Collections and Technical Services
ANSI	American National Standards Institute
ARL	Association of Research Libraries
ARPANET	US Government Advanced Research Projects Agency Network
ASCII	American National Standard Code for Information Interchange
BUILDER	Birmingham University Integrated Library Development and Electronic Resource
CDL	California Digital Library
CEDARS	CURL Examples in Digital Archives
CHIO	Cultural Heritage Information Online
CHLT	Cultural Heritage Language Technologies
CILIP	Chartered Institute of Library and Information Professionals
CIMI	Computer Interchange of Museum Information
CJK	Chinese/Japanese/Korean
CLIC	CLIC Consortium Electronic Journal project
CLIR	Council on Library and Information Resources
CNI	Coalition for Networked Information
CNRI	Corporation for National Research Initiatives
CORC	Cooperative Online Resource Catalog
CORDS	Copyright Office Electronic Registration, Recordation, and Deposit System
COUNTER	Counting Online Usage of NeTworked Electronic Resources
COVAX	Contemporary Culture Virtual Archive in XML
CSS	Cascading Style Sheet
DAVPPP	LC Digital Audio-Visual Preservation Prototyping Project
DCMI	Dublin Core Metadata Initiative
DCQ	Dublin Core Qualifiers
DDC	Dewey Decimal Classification

DEF	Denmark's Electronic Research Library
DESIRE	Development of a European Service for Information on Research and Education
DHTML	Dynamic HTML
DIGICULT	Digital Heritage and Cultural Content
DLF	Digital Library Federation
DLI	US NSF/DARPA/NASA Digital Libraries Initiative
DLS	University of Michigan Digital Library Services
DLPS	University of Michigan Digital Library Production Service
DLXS	University of Michigan Digital Library eXtension Service
DOI	Digital Object Identifier
DOM	Document Object Model
DOREMI	Document Management, Information Retrieval, and Text and Data Mining
DSSSL	Document Style Semantic and Specification Language
DTD	Document Type Definition
EAD	Encoded Archival Description
EASEL	Educator Access to Services in the Electronic Landscape
ECDL	European Conference on Digital Libraries
eLIB	UK Electronic Libraries Programme
ETC	Electronic Text Centre at the University of Virginia
FEDORA	Flexible and Extensible Digital Object and Repository Architecture
FGDC	Federal Geographic Data Committee
FRBR	Functional Requirements for Bibliographic Records
GILS	Government Information Locator Service
GIS	Geographic Information System
GML	Generalized Markup Language
HEFCE	Higher Education Funding Council for England
HTI	Humanities Text Initiative
HTML	HyperText Markup Language
IEMSR	JISC Information Environment Metadata Schema Registry
IFLA	International Federation of Library Associations and Institutions
ILMS	Integrated Library Management System
INDECS	Interoperability of Data in E-Commerce Systems
ISO	International Organization for Standardization
ITS	LC Information Technology Services
JCDL	Joint Conference on Digital Libraries
JISC	Joint Information Systems Committee
JPEG	Joint Photographic Experts Group

LC	The Library of Congress
LEADERS	Linking EAD to Electronically Retrievable Sources
MALVINE	Manuscripts and Letters via Integrated Networks in Europe
MARC	Machine-Readable Cataloging
MASTER	Manuscript Access through Standards for Electronic Records
MDA	Museum Documentation Association
METS	Metadata Encoding and Transmission Standard
MIX	NISO Metadata for Images in XML
MLA	Museums, Libraries and Archives Council
MOA	Making of America
MODELS	Moving to Distributed Environments for Library Services
MODS	Metadata Object Description Schema
MPEG	Moving Picture Experts Group
NCIP	NISO Circulation Interchange Protocol Standard
NDIIPP	National Digital Information and Infrastructure Preservation Program
NDLP	LC National Digital Library Program
NDMSO	LC Network Development and MARC Standards Office
NEH	National Endowment for the Humanities
NINCH	National Initiative for a Networked Cultural Heritage
NISO	National Information Standards Organization
NLM	National Library of Medicine
NPO	National Preservation Office
NSF	National Science Foundation
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information Systems
OASIS	Organization for the Advancement of Structured Information Standards
OCLC	Ohio Colleges Library Cooperative
ODBMS	Object Database Management System
ODRL	Open Digital Rights Language
OED	Oxford English Dictionary
OMR	DCMI Open Metadata Registry
ONIX	Online Information Exchange
OPAC	Online Public Access Catalogue
OSDLS	Open Source Digital Library System
OSS4LIB	Open Source Systems for Libraries
PDL	Perseus Digital Library

PEAK	Pricing Electronic Access to Knowledge
PERSEO	Personalized Multichannel Services for Advanced Multimedia Stream Management
PICS	Platform for Internet Content Selection
PRIE	Program for Research on the Information Economy
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RLG	Research Libraries Group
RLIN	Research Libraries Information Network.
RSLP	Research Support Library Programme
SCONUL	Society of College, National and University Libraries
SGML	Standard Generalized Markup Language
SLAIS	School of Library, Archive and Information Studies
SMIL	Synchronized Multimedia Integration Language
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SVG	Scalable Vector Graphics
SWWS	International Semantic Web Working Symposium
TEI	Text Encoding Initiative
TEL	The European Library
TIFF	Tagged Image File Format
TREC	Text Retrieval Conference
TULIP	The University Licensing Program
UDC	Universal Decimal Classification
UDDI	Universal Description, Discovery and Integration
UIUC	University of Illinois at Urbana-Champaign
UKOLN	UK Office of Library and Information Network
UMDL	University of Michigan Digital Library
UMIST	University of Manchester Institute of Science and Technology
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VLE	Virtual Learning Environment
VRML	Virtual Reality Modeling Language
W3C	World Wide Web Consortium
WSDL	Web Services Description Language
X3D	Extensible 3D
XHTML	Extensible HTML

XML	Extensible Markup Language
XML-QL	XML- Query Language
XQuery	XML Query
XSL	Extensible Stylesheet Language
XSLT	XSL Transformations
ZING	Z39.50 International Next Generation

Part I

Chapter 1

Introduction

This thesis examines the impact of XML on digital libraries. In order to set the context for the research, this Chapter will provide a brief definition of digital libraries, an analysis of their function in relation to traditional libraries and a description of XML. In addition, it will discuss metadata. We will see how XML can enhance metadata.

1.1 Context and Description

Digital libraries

Libraries, archives and museums collect, organize and provide access to materials, both textual and non-textual, which document culture and creativity. The collections held in these organizations connect us to our ancestors and preserve our cultural and intellectual record for future generations. Potential users come from the widest possible range of disciplines, ages and cultures. Within the past decade, the World Wide Web and other developments in the online environment have proved able to erase the boundaries of time and place so as to enable users to access, from a distance, information that has hitherto been available only through print or other formats, temporary or permanent exhibits, or to individuals (often accredited scholars) with special permission to visit an institution and examine the materials on site. Overcoming these barriers by using digital libraries has enabled new and wider audiences to access resources and to construct their own contents for the purposes of research, education and enjoyment. However, further barriers could have been placed in the way of the working of these new digital systems, had not markup standards been developed and become universally accepted: these issues are discussed in Chapter 2.

Although digital library development issues identified in this thesis might be applicable to all types of digital libraries, I concentrate attention on those digital libraries established by the academic and research sectors. Not being commercial, they were more open to being the subject of this research and pose greater challenges because of their wider range of materials.

The definitions and purposes of digital libraries are discussed in Chapter 3 but, as a brief introductory statement, it is worth noting that the term "digital library" was defined by Vice-President Albert Gore as an aggregate, implying electronic access to many sources of digital information [Information Infrastructure Task Force, 1994]. The collection, organization and provision of access to materials, which is the role of the traditional library, are, in my view, complemented by the work of digital libraries. Since the predecessor of the Internet, the United States government Advanced Research Projects Agency NETwork (ARPANET), was established in 1969 [Feizabadi, 1998] and the World Wide Web was created in 1989 [Berners-Lee, 1989], new technologies have become widely available. This has given an added dimension to the kinds of materials which libraries, museums and archives can make available, and has revolutionized the ways in which such institutions are used and organized.

Some regard the World Wide Web as a huge digital library, albeit a badly organized one in the view of librarians, though many researchers coming from the computing science disciplines, particularly the usability studies area, have a tendency to take it as given that any digitized collection, however organized, is a digital library [Rauber and Tjoa, 2001]. In some contexts, a digital library is a collection of materials which, having existed in a traditional format, have been digitized for a number of different reasons, for example to make them more widely accessible or to enable a different method of using the materials. Increasingly, though, digital libraries will include materials that never existed in any other form; these materials are now referred to as "born-digital". Nevertheless, as time goes on, it is likely that digital libraries will be increasingly involved in digitizing pre-existing materials. This is especially true in the Humanities, where there are large corpora of printed materials, such as original texts from the past, which are the core of their disciplines. These will provide, if digitized, added value in terms of their increased accessibility to their audience and also provide additional possibilities and problems for preservation. These points are particularly important in the digital libraries I chose as case studies for my research. For this reason, this research has omitted the area of the born-digital and concentrated rather on questions relating to the digitization of pre-existing materials

Digital libraries may be restricted by being on an Intranet or even on one computer only, or on a compact disk held at a particular location. However, the Web-based digital library can go beyond the traditional library and provide direct, immediate location and access to library materials without the user having even to visit a library or use a traditional catalogue. Easier and faster access to the World Wide Web permits access to large amounts of materials useful to scholars and researchers, though this is not always easy to find. These new tools and

technologies have expanded areas of scholarship, as well as enhanced the means of conducting research. To put it another way, these systems are beginning to contribute much towards the development of the information age. The possibility exists for rich mixed-media formats of digital materials to be transformed and connected into applications for education, providing significant tools for teaching, learning and research in an academic environment.

With the digital library acting as an information depository, there are many challenges concerned with the description of objects and repositories, interoperability and collection management. The underlying value of digital resources to the users will depend upon the quality of the contents, the organization, the data management systems and the presentation of the data. These mainly depend on issues of metadata and interoperability. Digital libraries suffer without solving the problems. The challenges of metadata and interoperability are now being investigated in the context of the latest technology, the Extensible Markup Language (XML), and these initiatives are being developed in the infrastructure provided by the XML environment.

Greenstein and Thorin [2002] wrote about three ages of the digital library: the Young Digital Library, the Maturing Digital Library and the Adult Digital Library. The Young Digital Library is obsessed with defining its mission, with securing funding, determining who are the right people to lead it and own it and its position in the organization which is, willingly or unwillingly or even unknowingly, its host. The Maturing Digital Library begins to think about its users and improve its internal management, both its technical organization and its management structure. To do this, it looks more carefully at optimizing the systems architecture. The realization comes about that no digital library can expect to do everything it needs to for its users single-handed, so it begins to long for standards. At this stage, it thinks about future proofing. One feature which has posed a challenge for the development of digital libraries is the need to ensure that any digitization undertaken will not become obsolete with the resulting loss of all the effort that has been put into the preparation and digitization of the materials. So, many initiatives have not even started but are waiting for a lead. This challenge caused by the need for standards and recommendations is beginning, as we shall see, to lead implementers to consider XML seriously. This is the stage which has been reached by the more advanced digital libraries that I saw while doing my research. The standards and recommendations, including XML, are still under development and have a long way to go before the digital library reaches the adulthood characterized by Greenstein and Thorin.

XML

The Extensible Markup Language (XML), like its ancestor SGML (Standard Generalized Markup Language), is a form of descriptive markup. It has been developed under the auspices of the World Wide Web Consortium (W3C). The first version of XML was originally published as a Recommendation in February 1998. XML was created to overcome the limitations of SGML and HTML for Web delivery [Bray et al., 1998]. We shall see in Chapter 2 that XML was needed because HTML is a fixed set of tags; SGML provides an extensible tag set but lacks Web support for network delivery. For this reason, XML has been designed for ease of implementation and for interoperability with both SGML and HTML [Bray et al., 1998]. XML is a vendor-neutral, platform-independent interchange initiative with a structured document representation, which can bring advantages for digital library procedures and services. XML is intended to allow fine-granularity markup of content and structure. It is becoming the markup language which is often used to create, archive and disseminate digital information on the Web.

XML has rapidly gained popularity as a markup language for information, finding constituencies in both the document-centric and data-centric worlds. A variety of Web applications and industry initiatives have announced their support for XML. Since XML was derived from SGML and is a technology that enables data to be supported on the World Wide Web, with features in flexibility and extensibility, the digital library community has found XML influential. Felstead [2004] predicted in her survey of the literature on integrated library management systems (ILMSs) published between 1999 and 2003 that the future of ILMSs would be to be integrated with a Web services platform; this implies XML. As Carvalho and Cordeiro [2002] have indicated, the role of XML in a library information system is to be found at three major levels: as a language that enables bibliographic data to be moved between systems and enhances interoperability; as a language that enables the complex kinds of data found in library information systems to be validated according to the standard specifications of particular data formats which they use; and as a language that enables services they offer such as search and retrieval to be made available in an efficient and more flexible way.

Thus, library materials encoded in XML can be converted into different formats, manipulated so that particular points could be displayed or transformed to create different displays. This therefore leads to reusability, which is one of the key concerns of those involved in activities to digitize any quantities of data. Most of all, and this is something that alternative formats like Adobe Acrobat or word processing formats cannot so easily do, XML is a tool which enables the creation of "hypertext" documents containing links to other documents or to other elements of the same document, for example, between one word phrase or sentence in a document and the

equivalent in the same or another document which represents its translation. Furthermore, an XML-based infrastructure can provide a flexible and platform-independent system for giving access to the materials in the digital library released from the constraints of proprietary library systems.

Paepcke et al. [1998] pointed out that interoperability would be a central concern among the world's scattered digital libraries. Digital library systems dealing with materials which are based on standard metadata schemas and encoded with the XML format are one of the ways of providing integration and interoperability between heterogeneous systems. An XML-based platform has the ability to become the platform of choice for sharing information because XML brings with it a large amount of independence from data formats, document models, and languages (or scripts), which Paepcke et al. thought was the long-term goal for digital libraries. This thesis demonstrates throughout how the digital library can use XML to build new collections, and to migrate legacy information or documents created in native format into XML for better persistence and enhanced usability such as that used in electronic publishing. Also, the digital library can use XML-aware databases to store and maintain materials to create library services.

Digital library projects have been actively developed around the world, as the Web technologies quickly advanced. Libraries have taken this path because the networked environment has made it possible. We can see by the increasing use of digital libraries that library managers believe that they can provide a distributed information environment that will allow people global access to many areas of information. Other important factors are that digital libraries are more convenient for users than conventional libraries, can allow multiple access to individual resources and can use technology to achieve the mission of long-term access and preservation. Currently, some digital library projects, such as my three case studies, are good models, having integrity and high quality. Some already play a vital role in the library community. Still others may be merely demonstrations, proof-of-concept applications. However, the encouraging movement of XML-based digital library development illustrates that XML rather than SGML is the best solution. Freter [1998b] remarked that any trade show demonstrates the popularity of XML among software vendors, and we ourselves could see this at library exhibitions such as LIBEX 2004 and Online 2004. DeRose [1999b] predicted that the computing industry would move to embrace XML. Although the XML family of specifications is still under development, this move to XML is a continuation of the trend that swept us all into an HTML universe; therefore, the role of XML in digital library development will only be increasingly influential with time.

Before the birth of XML, SGML had been accepted as the standard for document representation and thus had been adopted in many digital library projects, such as my three case study libraries. XML provides many of SGML's benefits not available in HTML, such as extensibility, flexibility and validation, and yet it is easier to learn and use than the complex SGML. XML is becoming much more widely adopted in the commercial world than SGML ever was: the reason for this is that commerce is embracing Internet 'publishing' for so many of its activities, and in this context it sees XML as an invaluable tool, where SGML had been well known only within the confines of publishing. Therefore, with such extensive industry support, XML is maturing and evolving to a family of specifications; it promises a more robust and flexible infrastructure for Web applications. Strong evidence from the standards Website at the Library of Congress indicates that standards made for the library and information community are transferring from SGML to XML, and new XML-aware initiatives are currently in process [Library of Congress, 2005b].

As early as 1998, XML was fast becoming the key language for an increasing number of Web applications [Sall, 1998]. As of 2005, it is poised fundamentally to change the way information is delivered and used as well as to enable the creation of new and powerful applications. Undoubtedly, digital libraries will be an important application for XML. Those sectors of the industrial world that used SGML are abandoning it and are embracing XML instead [NINCH, 2002]. Indeed, those digital libraries that do not use XML will suffer through the limitation of SGML discussed in the above sections. However, there is a need to explore how exactly the XML technology can help in digital library development. It would be appropriate to examine these issues in the context of practical large-scale digital library initiatives.

Metadata

Taking metadata in its widest sense, it is the basic material for library information retrieval. Catalogue records and data in abstracting and indexing services are traditional types of metadata. Day [1997] stated that when the library and information community discusses metadata, the most common analogy given is the library catalogue record. In a digital library environment, the effective management of networked digital information increasingly relies on the effective development and use of systems that can collect and use appropriate metadata [Day, 1999]. The diversity of formats of metadata reflects the existence of numerous subject communities, which in turn raises issues of interoperability.

In answer to these issues, a number of initiatives are emerging to provide much of the basic architecture for Web-based data. Firstly, Resource Description Framework (RDF) from the W3C

is a Recommendation for supporting the exchange of metadata on the Web. RDF uses XML as a syntax language and provides the infrastructure to facilitate modular interoperability among different metadata element sets that greatly support the integration of heterogeneous resources into digital library collections. Secondly, the Metadata Encoding and Transmission Standard (METS), also based on XML, was created to provide solutions for digital media with sufficient storage capacity, scalability, security, access, preservation and copyright provision to ensure a protection mechanism in a digital environment. As I learnt from my research interviews, the Library of Congress is implementing METS in its central repository and storage system. Many digital libraries are investing research efforts on the METS initiative and exploring how it works alongside metadata standards and the XML family specifications such as the XML intelligent linking ability to leverage its strengths [Library of Congress, 2005a].

A document structured by XML tagging will usually have well-identifiable elements of metadata. In the digital library context that is important because good information retrieval tools are essential and XML tagged metadata is more easily retrievable from the Web. As in many other fields, XML is also becoming the markup language used for the text itself in Web publishing, which will be evident in this thesis. It can also provide the framework for the insertion of the non-text into text, or for non-text documents themselves.

These three components, digital libraries, XML and metadata, are the main topics in this thesis, and I will discuss them more fully in later chapters. The key findings of this research are that firstly, through the research case studies in Part two of this thesis, I demonstrate that XML will have a major impact on the aspects of metadata and interoperability in digital library development. As discussed in Section 4 of Chapter 3, metadata was seen by the Library of Congress as the key to resolving challenges in building the digital library of the 21st century; metadata and interoperability were pointed out by a staff member at the Library of Congress as the main components for building global networked digital libraries. Secondly, as I point out in Section 2 of Chapter 2, SGML has limitations for Web delivery and, unlike SGML, XML was specifically designed for use in a Web environment. Thirdly, as noted above, increasingly, standards made for the library and information community are transferring from SGML to XML or are being based from the outset on XML. Considering all these facts, I estimate that digital libraries will suffer if they do not seriously consider XML technology because digital libraries are activities based on Web applications. We will see in Part two of the thesis that my three case study digital library initiatives have recognized the value of XML in digital library development and are following a policy of incorporating the new technology into their infrastructures as time goes on.

1.2 Research Issues

The digital library provides a digital service environment that is a comprehensive and integrated collection of information resources within a networked online information space where no distinctions are made between information formats. We have observed that XML is increasingly recognized by a broad range of Internet-based applications. Although there is consensus that XML should have some role in the digital library [Sperberg-McQueen, 1998; Arms, C.R., 2000; Jørgensen, 2001; Banerjee, 2002; Felstead, 2004], it is not always clear exactly how XML can be used in this context; therefore, I explain its use in metadata and content definition, and in order to evaluate in what range of situations it would be useful, I analyze by means of case studies its actual potential uses and its relationship with existing technologies in practice.

Through my study of real-world practices, I discuss in further detail these digital library development issues and specifically where the problems and difficulties lie. Also, after the discussion of problems, I point to XML efforts that could possibly help, and examine in what aspects the three case studies have gradually adopted XML technology into their infrastructures.

Finally, I take the lessons learned from this research investigation and produce a number of suggestions for library and information communities that have already launched a digital library project or are planning to develop one in the hope that they could avoid unnecessary risks and have more chance to succeed.

1.2.1 Goal of this Thesis

The goal of this thesis is to examine the strength of XML technology and explore the role of XML implementation in digital libraries. The thesis is also based on studies of the full practice of the three digital library initiatives, which were chosen because of their size and because they were examples in the global digital library community. These studies included research interviews conducted during visits in September 2002 to the three projects and HTTP usage statistics between 2000 and 2002 contributed to me by the staff of the projects. An online user survey was planned for a four-week period in 2005. This had to be abandoned because two of the digital library projects refused to permit it on the grounds of political and technical difficulties and the third made no reply despite repeated requests. This will be discussed more fully in Methodology Section (Section 2) of Chapter 6.

The Library of Congress is taken as an example of a national library. Its mission is to make its resources available and useful to Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations. As part of its mission, it has established the National Digital Library Program (NDLP) [Library of Congress, 1998e].

The University of Michigan Digital Library Services (DLS) is representative of an academic library. In June 2003 the DLS was renamed Library Information Technology [University of Michigan, n.d.]. Because the case study was undertaken while it was still called Digital Library Service, that is the name used throughout this thesis. Its mission is to support a virtual learning environment and preserve campus-wide materials for long-term access.

Perseus Digital Library (PDL) is an instance of a research and development testbed [PDL, n.d.]. It represents the mission of one Humanities scholar to use technology to facilitate research and teaching in the Humanities.

PDL has moved from SGML to XML, which demonstrates technically the ability to transfer between the two environments; Michigan uses SGML in its text collections and XML in its image collections, and is committed to transferring text creation from SGML to XML in their next version of middleware [Powell, 2004]; the Library of Congress is still using SGML, but they see the future direction as considerably more use of XML in the area related to metadata, and there is also an expectation that they will create an XML version of the American Memory Document Type Definition (DTD). They have been monitoring closely XML related technologies and have been deploying them in subsequent work in digital library development; that is, for completely new projects, XML has been used where SGML had been used in the past [Arms, C.R., 2004]. The three case studies have demonstrated from real life the move from SGML towards an XML environment.

I formulated the following questions to help me explore the impact of XML technology in digital library development:

- What are the origins of the coding of electronic materials?
- What are the strengths and limitations of markup languages, SGML, HTML and XML, for the digital library?
- What is the role of SGML in early digital library development?
- How is the field of the digital library evolving?
- What is the role and current development of XML and its associated technologies in

forming the foundation of Web data?

- What are the implications of XML in Web publishing?
- How are XML documents built in an XML-aware database management system?
- How are XML documents retrieved in an XML-aware database management system?
- What are the functions of metadata in a networked environment?
- How do the current practices of XML and leading metadata schemes support digital collections?
- What can in-depth case studies tell us about the potential role of XML in managing digital libraries?

1.2.2 Outline Literature Survey

Digital library-related research, development and services have attracted increased interest and effort. An increasing number of scientific disciplines, including library science, have made a contribution and many scientific communities have joined the field. Core disciplines and application communities increasingly engage in discussions and cooperation, stimulating open debate and the exchange of experiences, ideas, approaches and results. This can be seen by the following conferences: the Joint Conference on Digital Libraries (JCDL), sponsored by the Association of Computing Machinery (ACM) and the IEEE Computer Society; the European Conference on Digital Libraries (ECDL), the major European forum focusing on digital libraries and associated technical, organizational and social issues; ACM Hypertext; and the International World Wide Web Conference, organized by the International World Wide Web Conference Committee (IW3C2). These conferences are conducted mainly by the computing community, and their publications are mainly from computing related departments or laboratories. These tend to relate to technical aspects of digital libraries of interest to computing specialists and do not examine issues of interest to librarians, such as the management issues of digital libraries, or even information science issues such as those relating to classification schemes and other practices which have been developed for traditional libraries but can be applied to digital libraries.

For dissertations and theses, I examined online databases for the last seven years (since the birth of XML in 1998 to 2005) for research work completed or in progress covering the United States, Canada, Asia and Europe. The databases included: *Index to Theses in Great Britain and Ireland, Dissertation Abstracts International, UMI ProQuest Digital Dissertations, Networked Digital Library of Theses and Dissertation,* and *Current Research in Britain.* The results showed no dissertations or theses with the main topics XML and digital libraries. If there were theses with

XML-related or digital libraries topics, they were research works mainly from Computer Science or Information Management with no connection with Library and Information Studies.

In addition, I did literature surveys and researched journal articles in online databases. The databases include Library and Information Science Abstracts (LISA) and ISI Web of Science. I also looked systematically through the electronic journals in this field, *Ariadne* and *D-Lib Magazine*. The results are similar to the above searches for dissertations and theses. Also, I searched FirstSearch WorldCat List of Records, and retrieved only two books with the subject 'XML in libraries'. Both books were published in the United States in the years 2002 and 2003. Basically, the two books [Tennant, 2002b; Miller and Clarke, 2003] are subjects more related to technical issues in the digital library environment with specific examples such as inter-library loan and cataloguing. They do not include the full range of digital library development; for example, Tennant, out of thirteen case studies, only includes two on using XML for collection building and does not discuss management aspects in the digital library.

The result of the review of the related literature indicates that there does not appear to be any completed research subject similar to the topic of this thesis in the past years or at present in the fields of Computer Science, Information Management or Library Information Studies.

1.2.3 Contributions

As was indicated from the two books mentioned in the previous section which are on the subject of XML applications in libraries [Tennant, 2002b; Miller and Clarke, 2003], the digital library community is currently experimenting with XML technology in building or even in rebuilding digital collections, systems interoperability, cataloguing and indexing, databases, and in data migration. In general, XML technology has not yet been well-recognized in digital libraries. More real life practice and research would be extremely useful for digital library development. This strengthens my belief in the positive value of this research. I believe that the primary contributions of this thesis are the following:

- I undertake a digital image library experiment, which reveals practical aspects of the role and function of XML in modern database-backed Websites with generic application and I demonstrate that XML coupled with a database gives greater power than the sum of the parts in a Web application.

- I explore the potential of XML technology in digital library development and point out that the critical features of XML in building digital libraries are in metadata and interoperability.
- I make comparisons between the three digital libraries rather than merely investigating them individually.
- Although these outstanding digital library projects and others have been well-documented in journal articles, this research has brought information together into one place to give a bird's eye view of the entire state of the art and will contribute towards the success of newer projects which wish to take these findings into account.
- This research contains detailed analysis of a range of issues found in digital library development, including critical aspects that contributed to their success or failure, which they would not specify themselves.
- It is hoped that this research may be an inspiration to new digital library projects: from the research, I discovered that even the best examples of the digital library have encountered challenges.
- It is hoped that, in this thesis, digital libraries under development will be provided with information that could enable them to avoid the mistakes of others and to take advantage of the factors that have assured others' successes.

1.3 Outline of this Thesis

The research for this thesis is divided into two parts. The first part, Chapters 1-5, is organized as follows. After the introduction, Chapter 2 reviews the history of markup, the evolution of markup language from GML (Generalized Markup Language), SGML, HTML to XML and the technologies and standards of markup languages. This Chapter also reviews the history of SGML in digital publishing and shows how it laid the ground for the more universal XML. Furthermore, Chapter 2 also examines several efforts in electronic delivery of content, including audio and video content expressed in XML. Chapter 3 describes the elements of a digital library, including the evolving background, the collection and the functions of metadata in a digital library. Chapter 4 examines XML as a key technology in database management systems, including the concept of structured document and semi-structured data, XML data models in the relational and objective databases and XML query languages. Chapter 5 reviews numerous discipline-specific metadata formats represented in XML with reference to a number of existing projects and initiatives and also to the metadata interoperability infrastructure.

The second part, Chapters 6-10, is based on a research investigation into three well-known digital library initiatives all in the United States. These three digital library initiatives afford the opportunities for the hosting organizations, along with other agencies and organizations interested in collaborating with them, to mount research and development projects directed to real problems found in digital library development. These have included such issues as the selection of material for collections, digitizing, management of digital libraries, behavioural aspects of users and legal and economic issues.

This part is organized as follows. Chapter 6 describes the methodologies and gives background information to the case studies. Chapter 7 evaluates the rationale which the case studies gave for developing digital libraries, Chapter 8 evaluates the different digitization processes that have been used, and evaluates the metadata system and delivery methods. Chapter 9 discusses the management and organizational aspects, illustrated by comparing the case studies, good and bad practice, and mentioning the main innovations that the libraries in the case studies have made. Finally, Chapter 10 summarizes the research and finishes with a number of conclusions that, it is hoped, may benefit the development of digital libraries.

Chapter 2

The Evolution of Markup Languages

This Chapter reviews those full-text document technologies which preceded XML, including the history of markup, the evolution of markup language from GML, SGML, HTML to XML, and the technologies and standards of markup languages. It also reviews the history of SGML in digital libraries. It discusses their strengths and weaknesses particularly in the context of the digital library. This Chapter also examines electronic delivery of content in XML technology. It is followed in Part Two by other examples which are in the case studies, on which more detailed investigations are made.

2.1 History of Markup

In the early days of the traditional publishing industry, the methods of handling documents suffered many limitations. The printed document resulted from compositor, typesetter and copy editor manually formatting the text presentational structures (layout, page objects and glyphs) and logical structures (topic-structures, editorial-structures and characters) [Jelliffe, 1998]. Authors could spend a considerable time on the appearance of documents ensuring that the documents had the correct levels of headings, italicization, emboldening and indentation without paying attention to the logical structure of the document which would entail extra work with no practical advantage. Publishing therefore relied on markup languages which were instructions to printers to make text bold or italic or to indent or justify or perform other actions on the text to conform to the presentation that readers expected. The traditional purpose of markup was therefore to annotate text for a printer to embellish in ways that were impossible to note in handwritten or typed text.

When computers began to be used for textual data processing rather than just for mathematical calculation, standards were established for the representation of the characters within the data. These standards were necessary not because data were transferred between computers but more for the inter-changeability of equipment, such as tape readers, keyboards and other input/output devices. These standards were based on standards for data entry between existing electronic devices such as the teletype or telex. These standards included ASCII, the American National

Standard Code for Information Interchange, developed in 1963 (also known as ISO/IEC 646:1991) [Searle, 2002], which established a common representation for the letters of the alphabet, numbers, punctuation and other symbols. Diacritic characters such as acute and grave without which French, to name only one language, cannot be written correctly, were not included as the systems were biased towards the English language. Word processors, such as Wordstar, and desktop publishing systems, such as those intended for the Apple Macintosh, were developed which incorporated these standards but which dealt with document formatting and fonts in their own proprietary way. Since the goal was the production of a printed document, this did not matter. However, when networking was introduced and data were transferred between systems, be it online or by sending data on tape or diskettes, it became necessary to develop common standards for representation since it would have been awkward to convert on the fly between numerous different methods of representing characters [Bergerud, 1987]. The main outcome of this was the development of markup languages for data processing systems. They would notably indicate to the computer how to display material on the screen, or to the computer printer how to print out with suitable indentations, bearing in mind that a printer might be remote from the computer driving it. It would also identify individual characteristics of the document, enabling the representation of typographical characteristics which would lead to the text being independent of the markup which had hitherto produced only typographical distinctions.

In 1969, Charles Goldfarb [2004] of IBM together with Edward Mosher and Raymond Lorie had invented the Generalized Markup Language (GML) as a means of allowing text editing, formatting and information retrieval subsystems to share documents. Instead of a simple markup scheme, however, GML introduced the concept of a formally-defined document type with an explicit nested element structure. After the completion of GML, which never achieved any extensive usage, Goldfarb continued his research on document structures and these were incorporated not into GML but into what later developed as SGML.

SGML embodied features such as standardized information formats and structures that can be easily delivered over a network [Lander, 1997]. Additionally, because of national and international interest in the interoperability of systems in this area, SGML was developed as an American National Standards Institute (ANSI) standard and later international standard known as ISO 8879:1986. This was done under the collaborative efforts of the Computer Languages for the Processing of Text Committee (established by ANSI) and the Graphic Communications Association GenCode committee [Turner, 1990]. SGML provides a standard method of representing the electronic information contained in a document, independently of the system used for input, formatting, or output. With SGML, document publishing is largely improved in terms of document preparing, reuse and portability.

SGML is actually a meta-language for generating descriptive markup languages, giving permission for flexibility and customization [Coombs et al., 1987]. SGML allows all sorts of strategies, and thus is an ideal tool for large and long-term specific-domain industries such as scientific or military industries in dealing with ever-increasing amounts of data. SGML revolutionized the document-processing world, and quickly gained wide acceptance among major leading industries worldwide. Although SGML is powerful, its complexity critically limits it when delivering SGML over the Web [Arbortext, n.d.].

The accessibility of the World Wide Web and the global Internet has significantly changed the way documents can be stored, distributed and presented. Documents are being prepared less frequently with formatting information and more often with structural information by using tagging languages. With the tagging system, documents are encoded in the form of "tags", so that documents are readable without any formatting information and are easily interchangeable between platforms. Added to this, the tags can serve as metadata giving the possibility of querying information.

In 1989, Tim Berners-Lee of CERN (the European Laboratory for Particle Physics) developed a technology for enabling files on servers attached to the networks which made up the Internet to be viewed from computers with access to the network and thus the concept of browsing the Web was born [Berners-Lee,1989]. He based this methodology on that of SGML with its 'tags' but simplified it by giving it a fixed tag set and called it the HyperText Markup Language (HTML). The locations of documents on computers were defined by URLs (Universal, later Uniform, Resource Locators; now also complemented by URIs, Universal Resource Indicators) which were incorporated into HTML. This has allowed the development of an extensive Web of information, which has contributed to the wide popularity of HTML [Musciano and Kennedy, 1997]. HTML is simple, easy to link, with only one tag set for all applications. However, this limited tag set is an impediment to, amongst other things, extensibility, hierarchical structure and validation [Bosak, 1997].

A more flexible and expandable markup language, Extensible Markup Language (XML), has now been developed. It promises to solve the problems of diverse data types by allowing for user-defined markup rather than browser-defined markup. It allows markup that describes the content rather than the format. This description of content has implications for extracting and reusing the content in ways yet to be imagined.

XML inherits major features from SGML but without its complexity. XML goes beyond the limitations of HTML and positions itself as the solution for more demanding Web applications.

Williams et al. [2000] stated that XML permits data representation to be platform- and vendor-independent, thus allowing developers to describe and deliver rich and structured data from any application in a standard and consistent way. It would be a milestone for Web applications to see in the near future the widespread use of XML repository technology on a Web server, combining the best of both the database world and the XML world, which would change the Internet into a structured, interactive and easily navigable tool for our everyday use.

The evolution of markup languages is illustrated diagrammatically in Figure 1.

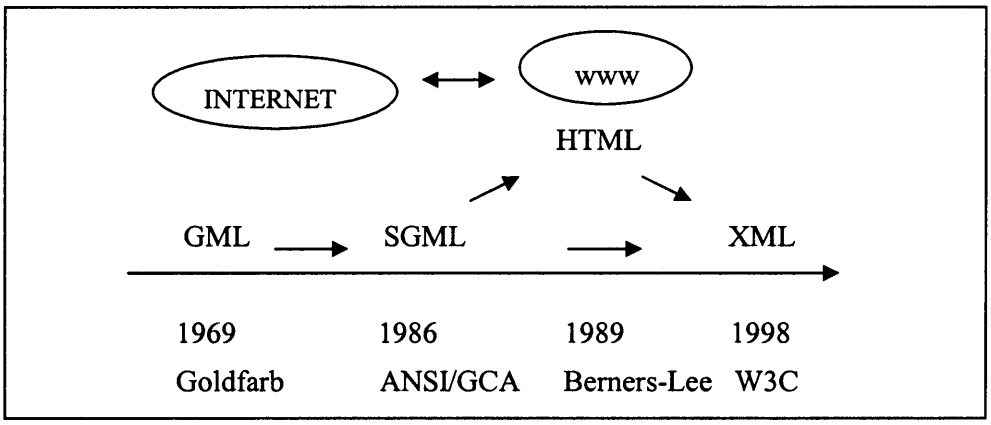

Figure 1: The evolution of markup languages

2.1.1 Types of Markup

As I have discussed in connection with traditional printing, so those text processing systems and the more recent word processing systems typically require additional information called "markup" to interpret for processing the raw text of the document. Markup serves two functions: it separates the logical levels of the document, and it specifies the processing functions to be performed on those elements [Goldfarb, 1990].

- **Procedural Markup** Traditional markup is known as "procedural markup" or "specific markup", which refers to a set of handwritten notations from the copy-editor. These notations cover instructions to a typesetter about how to lay out the copy and what typeface to use. With the progress of technology, batch-processing typesetting software allows the user to process large numbers of files using special codes that can be entered by using a computer keyboard such as nroff/troff and TEX [Coombs et al., 1987]. After a document is created, it can be processed by software programs to replace the embedded codes within the text of the document and generate proper layout. This so-called WYSIWYG (What-You-See-Is-What-You-Get) desktop publishing

software displays text and graphics on screen in the same way as they will print. Microsoft Word and WordPerfect are examples of this type of program.

• **Descriptive Markup** Descriptive markup, also known as "generic markup", describes the structure of the text in a document, such as a section or a paragraph, rather than its physical appearances on the page. Unlike procedural markup, the basic concept of descriptive markup is that the content of a document should remain separate from its appearance, which enables multiple presentations of the same information. The author can publish in various media and formats from one set of source files on paper, on-line and on the Web by applications. Examples of this type include SGML and its offspring XML.

2.2 Markup Languages

A markup language, as used in the sense of computer markup, is a set of markup conventions ("tags") used together for encoding texts and making them machine-understandable [Sperberg-McQueen and Burnard, 1999, Chapter 2]. In other words, markup language is a set of rules making sense to computers. It specifies what markup is allowed, what markup is required and what the markup means. Users define markup to make explicit an interpretation of a text. As long ago as 1990, Charles Goldfarb [1990] highlighted two main concepts:

1. Markup should describe a document's structure and other attributes rather than specify processing to be performed on it, as descriptive markup need be done only once and will suffice for all future processing.
2. Markup should be rigorous so that the techniques available for processing rigorously defined objects like programs and data bases can be used for processing documents as well.

In the following section, I describe SGML in detail because XML is descended from it. We can see many features that were retained in XML. We will also see the limitations of SGML which necessitated the creation of XML.

2.2.1 Standard Generalized Markup Language (SGML)

SGML is a meta-language; in other words, it provides a framework to construct various kinds of markup languages. SGML uses the principle of logical document markup, and applies this principle in the form of the definition of a generalized markup language.

2.2.1.1 Key Concepts in SGML

SGML separates the document into structure, content and style, but deals mainly with the relationship between structure and content [Bosak, 1997]. I will discuss this in the following sections.

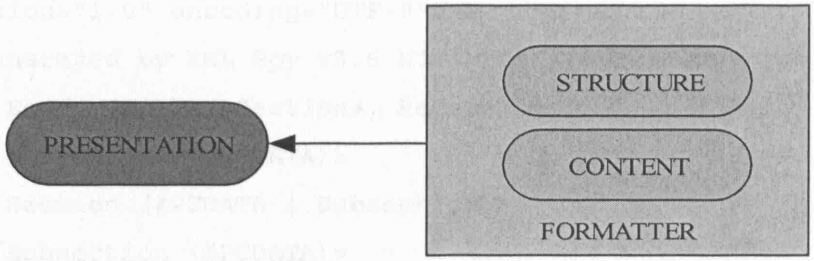

Figure 2: The SGML concept

2.2.1.1.1 Structure

SGML introduces the notion of a *document type* as its characteristics; that is, every SGML document is properly described by a Document Type Definition (DTD). The role of the DTD is to define a set of tags identifying all elements (the fundamental logical units of an SGML document) of a document and the rules expressing the relationship between the elements and its structure. These rules help to ensure that documents have a consistent and logical structure. Several document instances can belong to the same document "class". They are described by the same DTD, which means they share the same logical structure [Goossens and Saarela, 1995]. Figure 3 shows two instances of the same document class "essay".

Essay 1	Essay 2
Title	Title
Section 1	Section 1
Subsection 1.1	Subsection1.1
Subsection 1.2	Subsection1.2
Section 2	Section 2
Section 3	Subsection 2.1
Subsection 3.1	Section 3
References	References

Figure 3: Two instances with the same document class "essay"

A DTD names the elements that can be used; defines the content, indicates whether features are mandatory or repeatable or in what order they occur; defines possible *attributes* (attributes are used to provide additional information about elements) and their default values and defines the name of the *entities* that can be used. The DTD could be as follows using the example of Figure 3:

```
<?xml version="1.0" encoding="UTF-8"?>
<!--DTD generated by XML Spy v3.5 NT (http://www.xmlspy.com)-->
<!ELEMENT Essay (Title, Section+, References)>
<!ELEMENT References (#PCDATA)>
<!ELEMENT Section (#PCDATA | Subsection)*>
<!ELEMENT Subsection (#PCDATA)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT essay (Essay+)>
```

SGML also introduces the technique of *entity*, which means a named part of a marked up document. An entity might be a string of characters or a file such as <!ENTITY E4d "Information provision and electronic sources-XML">. Once the entity declarations have been made in the DTD, users can reference these in the documents using an entity reference with the form of &E4d. Parameter entities are distinguished from general entities by their use of a percent sign (%) rather than an ampersand (&). An entity may occur anywhere in the document. With entities, long documents can be made up of smaller files. Another advantage is that, should the content of the referred entity have been changed later, only the entity declaration needs to be changed as the entity reference will automatically call in the latest definition [Bryan, 1992].

2.2.1.1.2 Content

Content is the text itself and markup is information about the text held within an element or attribute. A tag is a piece of markup. One does the "tagging" to identify the content's position within the DTD structure. Taking the example of Figure 3, in the following short piece of structure, the tag <essay> indicates the beginning of the structure, and </essay> the end. The tag <essay> identifies a portion of a document and shows that "the following piece of structure is nested by the 'essay' tag". This is another characteristic of SGML that distinguishes it from other markup languages.

```
<Essay>
  <Title>Resource Description Framework</Title>
  <Section>Metadata on the Web</Section>
```

```
    <Section>Background</Section>
      ......
  </Essay>
```

2.2.1.1.3 Style

Since the major concept underlying SGML is to separate the representation of information structure and content from information processing specifications without regard to format or page layout, SGML defines no standardized formatting outputs. Therefore, other standards have been defined to specify formatting outputs. Document Style Semantic and Specification Language (DSSSL) (ISO/IEC 10179:1996) is an international standard designed as a transformation and style language for the processing and formatting of SGML documents [Bosak, 1996]. We will see in Section 2.3.3 of this Chapter that XSLT, which is the means for specifying formatting languages for XML, is a descendent of DSSSL incorporating many of its features.

2.2.1.2 Power of SGML

SGML was quickly adopted as a tool in information management by industries and various international leading organizations such as the European Union [Cover, 2002] because of the following powerful features [Arbortext, n.d.; Lander, 1997]:

- **Flexibility** SGML is extensive. With SGML, one can define and manipulate an information element at any level of detail. SGML was originally developed for large and long-term document publishing. It provides the capability for creating any desired set of tags, which is ideal for solving some of the most complex information processing problems.
- **Data independence** SGML is non-proprietary, thus the document can be managed within an Open Information Management (OIM) environment, in other words, a document is open to be processed by any program, not just by the program that created it.
- **Reusability** SGML allows the creation of different versions or formats of a document from a single file, which principally promotes efficiency and economy. The SGML document always remains usable and available. The user does not need to convert the document when hardware or software systems become obsolete because the files are ASCII files and this improves information longevity.
- **Economic characteristics** SGML separates the content and presentation, which allows document designers to focus on producing the document content. Designers can also improve efficiency by keeping only one copy of information so that they can produce many

documents for different purposes without the need to re-create the same information, thus increasing productivity.

Since SGML itself does not provide any formatting, specification of information processing, or semantic levels of information description, after became an international standard, work began on developing related standards to fulfil these functions. For example, Hypermedia/Time-based Structuring Language (HyTime) (ISO/IEC 10744:1997) provides element types to provide hyperlinking and other facilities [Goldfarb et al., 1997]. Such standards are more or less reflected in the development of the XML family of specifications, as we will see in Section 2.3.3 of this Chapter.

SGML was created as a solution for a broad range of database and document management (or publishing) problems. Most of the applications which have been developed for SGML have since been adapted for use with XML and are discussed more fully in Chapter 5 when illustrating XML-based applications. Also, in Section 2.1.4 of this Chapter, I will discuss the history of SGML before the advent of digital libraries, that is, the role of SGML in digital publishing.

2.2.1.3 SGML's Limitations for Web Delivery

Although SGML has been regarded as a useful tool, it has not generally during its lifetime been used in all applications. It has been restricted to the computerization of traditional publishing products and to selected new media such as databases on CD-ROM where proprietary structures have been encoded in SGML. Indeed, it has become less influential in the information processing domain because of the obstacles of:

- **Complexity** SGML has many complex options, which may only be used by small groups of users such as Text Encoding Initiative (TEI) used in Humanities and linguistics. These powerful features make it difficult to develop software due to the high-level SGML markup language. This defect further limits SGML data interchange because of varying levels of SGML compliance among SGML software packages.
- **Lack of vendor support** For a standard like SGML to become widely used, it needs the active support of browser developers. The two leading browser developers, Microsoft and Netscape, provided their full support to HTML instead of SGML because of the high expense of SGML software engineering, which is a significant barrier to SGML's widespread adoption. SoftQuad Panorama was developed as a standalone tool or as an extension application for Netscape [Lander, 1997]. Users were able to use Panorama to browse SGML in its native form. It provided powerful searching, broad presentation and style features, and enhanced linking

capabilities. But, Netscape supported it only for a short period of time, and it never became well known or understood by the majority of Web users.

• **Lack of Web support for network delivery** SGML uses content-based markup rather than presentation-based markup. The style process is a separate function which would be carried out by another program designed according to the specifications of DSSSL. Browser vendors have taken a short cut to avoid having to develop these special programs and have used HTML instead. Without the support of a browser and style-sheet, SGML information cannot be displayed in a readable form over the Web. Therefore, SGML files must be downward converted from SGML to HTML. This causes a certain amount of information loss [Arbortext, n.d.]. The lack of support for SGML in current browsers and the limitation of HTML in Web delivery, which I will discuss in Section 2.2 of this Chapter, mean that the advantages of SGML cannot be realized for the majority of end users of the World Wide Web.

DeRose [1997] described the early efforts on the subset of SGML, and pointed out that SGML software implementers did not make the most of the features provided in the software, and it was commonly found that the SGML parser did not cover all the SGML features. DeRose also pointed out that many SGML experts had proposed a formal subset of SGML. Bryan [1997] also thought the adoption of a smaller set of markup features should overcome the limitation of SGML in integrating with the Internet. XML is designed to be much easier to deliver on the Internet than SGML has proved to be, and much easier for software developers to implement than SGML in digital libraries. Since XML is a subset of SGML and it reduces complexity without reducing functionality, the movement to XML is expected to be substantial.

2.2.1.4 SGML in Digital Publishing

SGML predated the establishment of digital libraries because SGML helped to bring digital publishing into the digital age. We did not speak of digital libraries when SGML was in use in its early days because we did not have the World Wide Web to make the digital library a reality. Moreover, most projects were very firmly in the textual field since advances in multimedia had not yet come about.

The main quality of SGML is that it provides a rigorous procedure for defining the structure of a document, that is, by means of the DTD. Therefore, many major publishers such as Elsevier have adopted SGML for those documents with a complex formal structure which has to be retained even though the document goes through many versions such as formal publishing and industrial technical manuals [Tuck, 1996]. Two very early SGML applications, the Electronic Manuscript Project of the Association of American Publishers (AAP), and the documentation component of

the Computer-aided Acquisition and Logistic Support (CALS) initiative of the US Department of Defense, were examples [SGML Users' Group, 1990]. Applications in scholarly texts, such as the Association for Computing in the Humanities sponsored Text Encoding Initiative (TEI) Guidelines, also indicate that SGML could preserve the data over time in an independent and consistent format for access to the wider scholarly community for research purposes. With the emergence of the World Wide Web, large quantities of the world's research information are organized by using the SGML format and delivered electronically to support research, teaching and learning. This gave a good start to the early development of digital libraries. My three case study digital libraries are instances of this. In the following sections, I examine the history of SGML in academic digital publishing by dividing the applications into three areas followed by various projects worldwide: applications in language literature, applications in document management and applications in document preparation.

• **Applications in Language and Literature** The Electronic Text Centre (ETC) at the University of Virginia was regarded as the earliest worldwide endeavour in academic digital library development [Ream, 1993]. For most digital library or digital publishing with SGML content, the texts are encoded not in HTML but in more descriptive tag sets, such as the TEI Guidelines. HTML versions are then created for Web delivery. TEI provides a methodology which is ideal for the ETC. Each new work that is tagged can be added to an existing database of related works and can be searched and analyzed either as an individual entity or as part of a larger context. ETC uses the various forms of SGML markup to publish a variety of documents [Seaman, 1994].

The Lingua Parallel Concordancing project, on the other hand, was a project that digitized texts for linguistics purposes. Produced with EU funding, it was a corpus of texts and translations intended, among other things, to enable students to explore words in context in both the original and in translation [King and Woolls, n.d.]. This kind of concordancer had been developed before but only in a monolingual environment. In this context, the great variety of text types and conventions required control and it was decided to use the TEI Guidelines incorporating SGML to differentiate between whether a language was translation source or target, to group the texts into genres and to acknowledge copyright [Romary et al., n.d.].

• **Applications in Document Management** The application of SGML as the basis of a wide range of document management research projects has been the main theme at the DocMan Research Group, Department of Computer Science at the University of Helsinki [Linden and Heinonen, 1997]. Their research projects covered not only techniques in developing systems for processing structured documents but also tools to make the manipulation of documents easier. This included search tools for structured documents, a document transformation generator, an

SGML document manipulation language and the like. From an email communication, I learnt that DocMan has been renamed to DOREMI [DOREMI, 2004] and is still active, but the core technologies have been replaced by XML and its associated technologies [Linden, 2004]. For instance, one of DocMan projects, TranSID, which was mainly used for SGML transformations, was finished and there was no further development. For new projects, XSLT is used instead.

- **Applications in Document Preparation** Applications of SGML in document preparation exist for different kinds of materials. An early example of electronic journals in science is *Chemical Communications,* which is a parallel electronic journal. The CLIC Consortium Electronic Journal project, an electronic journal project from eLib, worked with the Royal Society of Chemistry to build an enhanced online version of *Chemical Communications*, which substantially influenced the standards for electronic information within the field of chemistry. CLIC experimented with new formats such as Virtual Reality Modeling Language (VRML), displaying the Chemistry Markup Language (CML) within SGML structures with specially-developed viewer, which was equivalent to a Web browser [CLIC, 1996].

SGML has been found valuable for use in digitizing manuscripts. Chaucer's *Canterbury Tales* is one of the earliest works in English literature. Scholars are trying to decide on the form of the original text, based on the copies we have now, using techniques of textual criticism [Robinson and Solopova, n.d.]. Additionally, to help to determine the authenticity of any alternative texts, it is necessary to make a complete concordance and a grammatical analysis of the texts and of any references to the texts made by medieval scholars. The different versions need to be linked to each other, and to any images. The texts have all been transcribed into plain text, and then a program called Collate is used to convert the text into Collate encoding which is a simpler form of markup than SGML [Robinson and Solopova, n.d.]. A CD-ROM was produced, using a form of SGML which was developed from the Collate markup. This has application outside the realm of Chaucerian scholarship since philologists are interested in the language, and the deep analysis of the text has resulted in a large amount of data. All this additional data, in some cases interpretation, has been made available through the marked up text in a way that would not have been possible without markup such as SGML which was used in the Collate variant as that was the best tool available at the time.

With the advent of the digital age, research methods for scholars have been radically transformed from paper to digital, and data processing has opened up numerous opportunities for lexicographical work over the years. A notable case of this is the *Oxford English Dictionary (OED)*. The Dictionary was made available as an online publication in March 2000 designed to take full advantage of the medium of the Internet [New, 2000]. The *OED* Design and Production team provided data marked up in novel SGML encoding scheme. The HighWire Press built a

system in Java which used the Verity search engine to search the Dictionary [New, 2000]. According to an email communication, the OED is in the process of transforming its editorial database to XML [Warburton, 2004]. By 2006, all data on the OED Online host will be in XML, but before that time, while the editorial system is in two metalanguages, the XML will be converted back to SGML for the online host.

SGML has been around since the middle 1980s and has remained quite stable as an industry-wide standard through the fact that it provides data longevity, portability, and even paperless publishing and distribution. From my research, I noted that SGML has proven effective as a document creation, management and delivery solution, which has laid a good background for the later development of XML in digital libraries. Much of this stability stems from the fact that the language is both feature-rich and flexible. This flexibility, however, comes at a price, and that price is a level of complexity that has inhibited its adoption in a diversity of environments, including the World Wide Web. Perhaps because of this, some projects have come to an end, some are still working with SGML, and some have continued to grow and are changing their platforms to the new technology, XML.

2.2.2 HyperText Markup Language (HTML)

We have seen that SGML is an excellent format in which to store data, while its size and complexity have proved to be difficult on network delivery. What was needed therefore was a system with future-proofing, supported on many platforms, extensible to multiple data formats in order to overcome the limitations of SGML. The HyperText Markup Language (HTML) which can be defined as an SGML markup language, an application of SGML, is the most successful document format developed so far [Arbortext, n.d.]. It is simple, portable, has built-in stylesheets, and can be created and processed by a wide range of tools, from simple plain text editors to sophisticated WYSIWYG authoring tools. HTML is currently the building blocks of the Web [Maler, n.d.].

Although the strength of HTML has been widely recognized in most Web applications including digital library applications, the limitations of HTML in Web applications are also well-known and these are the barrier to integrating more Web applications [Arbortext, n.d.; Bosak, 1997; Lander, 1997; Sall, 1998]. For example, HTML does not provide a mechanism for checking the data for structural validity or imposing editorial control. This is more important in a multimedia digital library environment than in the average commercial application because digital libraries by their very nature (being documents produced by many different authors to convey any information they

wish within the universe of knowledge) are less structured and controlled than commercial applications which restrict their data content to a limited range of transactions.

The limitations have only been overcome through the development of a further flexible, extensible and standardized Web technology, XML.

2.2.3 The XML Effort: "SGML on the Web"

XML is a common syntax for expressing structure in data. It is defined by the World Wide Web Consortium (W3C). As of 2005, the current version is XML 1.0 Third Edition, W3C Recommendation 4 February 2004 [Bray et al., 2004]. XML was originally developed for the Web. The group of experts who created XML was very familiar with SGML syntax [Bray et al., 1998]. They designed XML as a solution to the automation of the Web that neither SGML nor HTML could achieve. Its main purpose is indicated in the XML specification as being to make it easy and straightforward to use SGML on the Web, easy to define document type, easy to author and manage SGML-defined documents, and easy to transmit and share them across the Web [Bray et al., 1998]. Like SGML, XML is a meta-language. It can be used to create domain-specific vocabularies such as mathematical, chemical, and music markup languages. XML goes beyond the limitations of HTML; that is, it allows fully extensible, validation supports, reusable information and internationalization which are key features needed for a new generation of "weblication" (Web application) [Freter, 1998a].

2.2.3.1 Valid and Well-Formed XML Documents

There are two types of XML documents: valid and well-formed documents. An XML document with correct syntax is a well-formed XML document, and does not require validation against a DTD. In general, if one is maintaining a set of documents, then a DTD is advisable. However, if one is creating a single document, a well-formed document is good enough to be read by an XML processor which, like a Web browser, can read the XML data and display them. Although a well-formed document may be DTDless, it should, however, be well-organized and meet some rules [Bray et al., 1998]. For example, a document must contain one root element; elements must nest inside each other properly; all attribute values must be enclosed in quotation marks such as a *div* element with the attribute *class* having the value *preface*: .

2.2.3.2 Relationship between SGML, XML and HTML

Like SGML, XML is a meta-language providing a mechanism to specify new languages. Unlike

HTML, XML lets users define, adjust and grow their markup languages to any extent desired, while HTML defines a way to describe information in only one specific document model. SGML is the "mother language" and, of course, predated the World Wide Web. XML being a systems application of SGML, XML documents are conforming SGML documents [Bray et al., 1998]. An XML document does not necessarily have a DTD, while every SGML document must conform to a DTD, and that makes the main difference in network delivery between SGML and XML. SGML instances require a DTD to be available, while XML instances need only to be well-formed. HTML is simple and was created to view the data; SGML is big, complex and is ideal for large-scale, highly structured document management. XML is extensible and is suited to handle data semantics [Sperberg-McQueen, 1998]. HTML's one-way hypertext link has been viewed as relatively simple yet useful, whereas XML provides more powerful new features linking capability [Maler, n.d.]. XML is not created to replace HTML; rather the two can be regarded as complementary to each other [Sall, 1998]. HTML is only processible through a Web browser; XML may also be so accessible but can also be processed outside the Web environment.

2.2.3.3 XML-Related Initiatives

The W3C XML Working Group controls most of the specifications relating to XML. Some are official Recommendations, and others are still under development. In the following paragraphs, I introduce several initiatives that form an essential component of the entire XML family:

- **XML Pointer Language: XPointer Framework, XPointer element() scheme and XPointer xmlns() scheme (W3C Rec 25 March 2003); XPointer xpointer() scheme (W3C Working Draft 19 December 2002), XML Base (W3C Rec 27 June 2001) and XML Linking Language (W3C Rec 27 June 2001)** XML Linking defines a standardized, nonproprietary method to represent links among resources. XLink adopts the HTML concept of one-directional "simple links". Further, XLink supports more powerful "extended links" with which the user could create many links in one or more XML documents. XLink's indirect links largely help to eliminate the familiar "broken hyperlink" as exists in today's Web architecture [Maler, n.d.]. Once a Web page address changes, the author would be able to update all the links that point to it by editing just one file instead of every file that points to that URL.

XPointer is built on top of the XML Path Language (XPath, W3C Rec 16 November 1999), which is an expression language underlying the XSL language [DeRose et al., 2001]. In HTML, it is possible to link to the middle of a page only if the author of that page puts an anchor tag there. XPointer takes advantage of XML's tree structure, thus one would be able to "address to" any element tree of an XML document. For example, the expression child (2, course).child (3, book)

locates the third child element of book of the second course in the XML document.

XML Pointer borrows concepts from HyTime and the TEI extended pointers [DeRose et al., 2001], of which HyTime defines link structures and some semantic features and TEI provides structures for creating links, aggregate objects and link collections.

• **Extensible Stylesheet Language (XSL) Version 1.1 (W3C Working Draft 28 July 2005)** XSL is the mechanism to specify how the XML document should be displayed. Although XML supports Cascading Style Sheet (CSS) for style and page-layout issues, XSL is far more powerful than CSS. CSS works by annotating existing document structures and can be used to style HTML and XML documents; XSL transforms the existing document into a new document tree which consists of formatting objects [Bos, 2005]. For example, XSL can be used to transform XML data into HTML/CSS documents on the Web server; thus, the two stylesheet languages, XSL and CSS, can be used together.

XSL uses XML grammar, and large parts of it are inherited from DSSSL [Bosak, 1997]. XSL is actually a family of three Recommendations produced by the W3C's XSL Working Group: a language for transforming XML documents (XSLT), a language for defining XML parts or patterns used in the transformation (XPath), and vocabulary for specifying formatting semantics (XSL-FO). XSLT is the most important part of the XSL standard; it can be used to transform an XML document into a format like HTML that is recognizable to all browsers. This enables XSL to have power by working with HTML and server-side processing tools on the Web such as scripting language [Bos, 2005]. I will discuss some examples of XSL in Section 7 of Chapter 4.

• **Namespaces in XML (W3C Rec 14 January 1999)** The XML Namespace Recommendation takes care of the expressing of universal names in a processing application [Bray et al., 1999]. When two different documents use the same names describing two different types of elements, we must specify which tag set the element comes from to prevent potential conflicts. For instance, the <figure> tag could be the representation of person, numerical symbol or number, or diagram. Using the namespaces, developers can use the vocabularies from different domains without causing conflicts.

• **XML Query (XQuery) (W3C Working Draft 4 April 2005)** XML Query is designed to provide flexible query facilities to extract data from real and virtual documents on the Web [Malhotra et al., 2003]. Therefore, XML Query can be implemented in many environments like traditional databases, XML repositories, XML programming libraries and so forth. Queries may combine data from many sources. For instance, with similarities to SQL, the language has an SQL-like SELECT-WHERE construct, and borrows features of query languages recently developed by the database research community for semi-structured data. XML Query borrows path expressions from XPath. For this reason, the XPath specification is also being revised by the

XML Query committee. The XML Query language is not complete yet; it is still a work in progress. Nonetheless, implementations of XML Query are already making their way to the market [Ivanov, 2003]. I will give some examples of XML Query syntax in Section 5 of Chapter 4.

- **Document Object Model (DOM) Level three (W3C Rec 7 April 2004)** DOM specification defines a common application programming interface (API) for accessing HTML and XML documents from a Web browser. DOM level 1 and 2 reached Recommendations status in the year 2000, while DOM level 3 Core Specification dates from year 2004. DOM is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content and structure of documents on the fly [W3C DOM Working Group, 2005]. In Chapter 4, Section 7, I present some examples of manipulating XML documents with scripts via DOM.
- **XHTML (W3C Rec 26 January 2000)** XHTML 1.0 combines the strength of HTML with the power of XML 1.0, reformulating HTML as an XML application. One of the main missions of XHTML is to develop new XHTML-conforming modules which allow the design of new user agents across various platforms such as the browser-based phone [W3C HTML Working Group, 2002]. Developers migrating their content from HTML to XHTML 1.0 will not only benefit from the XML-related technologies but also assure their future compatability. XHTML is the successor of HTML, and a series of specifications has been developed for XHTML, with some output from W3C's earlier work on HTML 4.
- **XML Schema (W3C Rec 2 May 2001)** XML Schema describes the structure of an XML document. It is a superset of SGML DTD. However, the XML Schema can be more specific in defining and describing an XML document.

Unlike DTDs, XML Schemas are written in XML syntax. One of the features of the XML Schema is the rich data typing which enables a user to define data by type such as character, integer, and many other elements. This approach is an advantage for data exchange among applications or databases. For example, a DTD might have a tag designated as <price>, but the content contained within that tag could be a number or a character string. A schema allows the definition of the document data type. The XML Schema provides benefits when the mechanism in a DTD may not be sufficient to describe a real world complex relationship for data exchange among applications or databases [Mertz, 2001]. DTD originally was designed for use with text. It provides few facilities in describing document contents; XML Schema, on the other hand, allows users to reuse content models, to specify the number of occurrences, and to name groups of elements. An XML Schema is written with XML grammars, and for this reason, it is extensible to further additions [Mertz, 2001].

The structure of an XML Schema is more repetitious than that of a DTD; tools for creating XML

Schemas or converting DTDs to schemas are available which I will discuss in the next section. The XML Schema poses a substantial high learning curve; the real world implementation limitations on XML Schema need to be addressed and discussed before it is widely accepted [Apache Software Foundation, 2000]. We have seen that the SGML DTD has contributed to structured information processing and data interchange and the XML DTD will continue to do so. I will discuss further the latest developments of XML Schema in Section 2.1.1.2 of Chapter 8.

- **Simple Object Access Protocol (SOAP) (W3C Rec 24 June 2003)** SOAP is a lightweight communication protocol drafted by Microsoft, IBM and others for accessing services on the Web [SOAP, 2003]. It is an XML based protocol with the same power of simplicity and extensibility as XML. A SOAP-based interface could possibly provide a simple and quality search on the Internet. Because of the simple exchange mechanism, SOAP could potentially be used in combination with a variety of other protocols. Yet, SOAP's XML protocol requirements still need wide public discussion before it becomes a mature technology [SOAP, 2003].

- **Web Services (W3C Working Draft 11 February 2004)** The Web Services Architecture specification gives the definition of Web Services as: "A Web Service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format. Other systems interact with the Web Service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards." [Booth et al., 2003]. In other words, Web Services enable applications to talk to each other in an open networked environment [Gardner, 2001]. Web Services have the potential to provide functionalities needed in an open and distributed digital library environment such as online public access catalogue (OPAC), authentication, charging mechanisms and any other discrete process required by the digital library. As of 2005, the Web Services family of specifications are still under development as W3C Working Drafts. However, discussions, research and industrial activities are currently very active as its massive potential in building Web applications for electronic commerce evolve [Chen et al., 2003]. I will discuss its potential further in Section 2.3.1 of Chapter 8.

2.2.3.4 XML-Aware Software

The mainstream software business world is developing fully-conformant XML software. Using these software packages, application and tool, developers could possibly build XML-enabled applications for e-commerce, application integration and Web publishing.

In this Section, I highlight software for XML with different categories: editors, middleware, server software and special-purpose software such as XML-aware library systems, Web

publishing tools whether it is freeware, shareware or a commercial product. But I do not cover database support here as this will be discussed in Chapter 4. I refer to several XML software resource lists including Web Developer's Virtual Library, XMLSoftware.com, XML Software Tools at OASIS, XML Buyer's Guide at O'Reilly, Tucows Software Library and CNET Software Library.

As far as a digital library is concerned, there are a number of tools available. The first requirement is to have a package which includes authoring/editing tools for the interactive creation, modification and composition of XML documents. These include the commercial product XML Spy, which could be the most complete XML set of tools on the market. It includes XML and XML Schema editing and validation, and XSL editing and transformation in one product. I learnt from research interviews that my case studies in Michigan and the Library of Congress have also been planning to deploy it.

A conversion tool is needed because many digital libraries are still using SGML, as those of our case studies of the Library of Congress and Michigan. These institutions have already invested heavily in SGML, which governs the structures and describes the contents of their documents. OmniMark is a programming language with built-in SGML and XML facilities that can undertake this conversion. I learnt from the research interviews that the Library of Congress used this free product in its American Women project. The MIMAS team, which I will discuss in Section 3.1.2 of this Chapter, used OmniMark as well.

An XML server can be an XML-aware Java 2 Platform, Enterprise (J2EE) server, a Web application server, an integration engine, or a custom server which contains an application development environment and may provide access to data in a variety of data stores [Bourret, 2004]. Tools in this area include AxKit, which is a free XML application server for the Apache Web server. It can operate by simply applying an XSL stylesheet to retrieve data from a database. The University of Saskatchewan Library found this product useful and was able to migrate their law document repositories from HTML to XML with little budget [Fichter, 2002].

There are a number of types of middleware which are general software packages used by making XML-aware applications transfer data between XML documents and databases. Application integration tools include development toolkits like Java Project X from the Sun XML Library, which is a set of core XML-enabling services written completely in the Java programming language. A large number of tools in this area are available as open source software.

There are a number of integrated library systems which use XML. Of these, Ex Libris have developed an information portal named MetaLib which provides libraries, institutions and consortia with a standardized user interface for managing information systems [Ex Libris, n.d.]. MetaLib is open architecture and supports initiatives such as MARC, Unicode, OpenURL and XML. My case study, the University of Michigan, has signed up to this library system and at the time of the case study hoped that this would lead to more integration between the library system and their digital library. I will discuss more on this in Section 2.1 of Chapter 8 and Section 3.1.3 of Chapter 9.

Content management and publishing tools are also needed for the digital library. These are tools for the electronic delivery and display of XML documents. Their purpose is to support the full document lifecycle. For example, Java-based Cocoon is an open source XML publishing framework that can be used to publish XML on the Web as HTML. It supports the use of DOM, XML and XSLT to provide Web content. The eScholarship initiative at the California Digital Library demonstrated that the combination of XML, XSLT and the Cocoon publishing framework could be achievable and effective to the project goal in building electronic scholarly print-based communication [Tennant, 2002a]. More tools in this area provide users with the ability to use their existing word processors such as Content@XML, which enables Microsoft Word and Adobe to provide XML content. Other tools provide an integrated solution for managing and publishing dynamic Web content such as Dynabase, which can also be personalized for end users.

As the collection of papers in Tennant [2002b] pointed out, digital librarians have been struggling to find economic and suitable XML tools to solve library problems or to create new opportunities. Tools that work out well in one project do not necessarily work in the same way in a different scenario with different supporting technologies. I suggest that the digital library community experiment and develop using these tools and disseminate their results so that other projects can learn from the experience of earlier projects.

2.2.3.5 XML and Multi-Script

I mentioned in Section 1 of this Chapter that ASCII had enabled computers to talk to each other. This was true only when they were dealing with the English language or other languages which use no diacritics. If it was necessary to transfer data containing any alphabetic characters outside the normal 26 found in the roman script then extended ASCII was used, and there were many extensions for different purposes. ASCII is a seven bit character set using the 8th bit for a check

digit. Extended ASCII uses 8 bits for the representation of characters offering 256 combinations [Searle, 2002]. This is not enough for languages using pictograms such as Chinese.

It was decided to develop a two-byte character set that is now known as Unicode (ISO/IEC 10646), which would offer enough combinations for all the characters from most of the world's scripts if not all [Unicode, Inc., 2005]. At the same time, it has been adopted for use with XML, with various conventions having been adopted for the representation of the characters within XML [Dürst and Freytag, 2003]. This of course adds to the universality of XML. The *Digital Dictionary of Buddhism (DDB)* is an example of such an application [Muller, 2001].

Unicode provides full support in interchange, processing, and display of the written texts of a wide range of modern world languages [Brown, 2001]. This mechanism made available XML/Unicode based data which could be searched and displayed in multiple scripts for all digital libraries accessible on the Internet [Erickson, 1997]. This technique has been used not only for dictionaries but for any kind of machine-readable publishing.

2.3 XML and Text-Based Content

2.3.1 Text-Based Content Technologies

2.3.1.1 Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is a set of generic guidelines for the representation of textual materials in electronic form enabling researchers in any discipline to interchange and reuse resources, independent of software, hardware and application [Burnard, 1995].

The TEI Guidelines defined four hundred SGML (now XML) elements and associated attributes, which can be combined to make many different DTDs, suitable for many different purposes, either simple or complex [Sperberg-McQueen and Burnard, 2002]. The various elements in TEI are grouped into tag sets. Some of these elements which constitute the TEI Header are metadata. In a sense, metadata can be regarded as part of the content in the same way as the title and author printed in a book are metadata and at the same time part of the content of the book. According to the Guidelines, the main TEI DTD is made up of combinations of smaller tag sets, which include:

- **Core tag sets** tag sets defining elements likely to be needed by all documents.
- **Base tag sets** basic components for specific text types, such as the encoder may choose the prose base for prose texts, and the verse base for verse.
- **Additional tag sets** extra tags useful for particular needs; the encoder may add them to the selected base.
- **Auxiliary tag sets** used for the encoding of ancillary descriptive information.

The TEI core consists of two core tag sets: a large number of elements available to all kinds of documents, and the Header, which is metadata which, as we will see in Section 2.2 of Chapter 5, supports the same function as an electronic title page for the electronic text. The tag sets are extensible to enable markup of new types of material. The EAD and the CIMI are two successful examples based on the TEI scheme which I will discuss in Section 2.3 and 2.4 of Chapter 5.

A TEI conformant document will have a Header preceding the text itself. The text consists of optional front matter followed by optional back matter. Within text, all the main TEI bases use a generic subdivision element with attribute *type*= (followed by, for example, "part" or "chapter") within the element <body>. This allows the users to define all needed subdivisions but without the necessity for a complex set of different tags, though can be numbered <div1>, <div2> and so forth.

The XML version of the TEI Guidelines (TEI P4) was completed in March 2002. The new version of TEI can be processed and maintained using readily available XML tools instead of the special-purpose ad hoc software originally used for TEI P3. Moving ahead, the TEI Council has highlighted several directions as following main strategies. Among them, the TEI P5 will not have the same compatibility constraints as P4, but rather have better articulated strategies for interacting with other standards [Sperberg-McQueen and Burnard, 2002]. P5 was scheduled to be released by the end of 2004. According to Rahtz, P5 is based upon Relax NG, which is an XML Schema language. P5 is expected to support multimedia and graphics, XML namespaces and other capabilities [Rahtz, 2004].

Sometimes, there is a lack of clear distinction between content and metadata: for example, in manuscripts where the original often has difficulties in transcription which have to be supported by markup which could sometimes be regarded as metadata. XML can help this lack of clarity. Manuscript Access through STandards for Electronic Records (MASTER) is an example of this. MASTER is a European Union funded project to create a single on-line catalogue of medieval manuscripts in European libraries [MASTER, 2001]. The development of the MASTER DTD for manuscript descriptions co-operated closely with the TEI Workgroup on manuscript descriptions.

It developed an agreed international standard for manuscript records, based on the TEI implementation SGML and XML. MASTER also developed software tools to ease the preparation of manuscript descriptions conformant to the standard, such as an online XML parser to validate against the MASTER XML DTD, an online XML viewer and so on.

2.3.1.2 Electronic Journals

In an academic environment, rapid access to research material by many users at the same time has become invaluable to researchers, particularly those in fast-changing science fields such as medicine where journals are the main sources of information. A recent evidence of this XML approach is the project of the National Center for Biotechnology Information (NCBI), a centre of the National Library of Medicine (NLM), which created in March 2003 the first version of the Journal Publishing DTD which was intended to provide a common format for the creation of journal content in XML [NLM, 2003]. The underlying idea of this DTD is to simplify the creation of journal content by using XML technology.

Electronic journals are now to be seen in every research field. *New Left Review (NLR)* is an example in social science. *NLR* applies TEI/XML techniques to manage its online archive. Journal articles were digitized in accordance with the TEI and edited in XMetal [TEI, 2001]. The initial XML files were then processed by XSLT stylesheets, which ultimately produced the HTML Web pages. Furthermore, *NLR* is planning to reuse the composite XML files for libraries, for example storing articles on CD-ROM.

Electronic journals were marked up in SGML before XML become available, as I discussed in Section 2.1.4 of this Chapter, under the role of SGML in digital publishing. The following is just one example of how one organization is approaching the problem that would be caused by a move from SGML to XML. The MIMAS team at University of Manchester thought that as electronic journals publishing moves towards being XML-based, there will be a requirement for retrospective conversion to XML of existing data currently held in SGML. They launched a project in the year 2000 to change the existing SGML-based electronic journals to the XML format [Apps and MacIntyre, 2000]. OmniMark was the core technology because it was SGML and XML aware. The project came to the conclusion that the significant disadvantage to using XML was the relative instability of the XML family of technologies because the XML environment is immature. Although the approach appeared years ago, it still reflects the situation of XML currently in electronic publishing. Nevertheless, these shortcomings of XML should be overcome as technologies stabilise.

It seems to me that a requirement to use a standard interoperable format for electronic publishing will become increasingly important as academic publishing moves towards becoming globally accessible through the Web linking mechanism implied by initiatives such as CrossRef discussed in Section 1 of Chapter 5. Ideally, XML technology allows the freedom to publishers of making the parameters of their documents self-defining, by providing functions which permit much more database-like operation. Thus, by using XML, publishers will have much more flexibility to design the appearance, types of links and embedded programs. Furthermore, the increased functionality of XML will also allow the possibility for more complete interoperability with popular word-processing software. As we have seen in Section 2.3.4 of this Chapter, there are several XML-based products available on the market, either open source or commercial products, which provide an electronic publishing environment including workflow management.

2.4 XML and Non-Text Content

2.4.1 Non-Text Content Technologies

The format of multimedia data requires successful interaction of a complex system of hardware and software. Failure of any part of that system means that it could not easily be used by its intended audience, nor could it be migrated to new data formats, and so would become inaccessible due to obsolescence of technology. Hence, their representations are much more challenging than text data, and a standard format for them matters crucially in order to keep them over time without considering the obsolescence of the technology. As of 2005, there are more proprietary data formats than standard formats in the multimedia community, for example, Windows Media Audio from Microsoft, QuickTime from Apple and so on. Below I examine an XML-based standard MPEG, which is used in one of my case studies, the Library of Congress.

2.4.1.1 Moving Picture Experts Group (MPEG)

The Multimedia Content Description Interface (MPEG-7) became a formal standard (ISO/IEC: 15938) in 2002 [Martínez et al., 2002]. It was developed by the Moving Picture Experts Group (MPEG), a committee of the International Organization for Standardization that focuses on the encoding and processing of motion pictures, audio, and related multimedia formats. MPEG-7 is expected to provide standardized description language and schemes for concise and unambiguous content description of data/documents of complex media types [MPEG, n.d.].

The MPEG-7 community is attempting to combine efforts with other standard groups through liaisons such as EAD, Dublin Core and W3C. At the same time, it adopted XML Schema as the MPEG-7 Description Definition Language (DDL), which defines the syntactic rules to express and combine Description Schemes and Descriptors [MPEG-7 DDL, n.d.].

The MPEG-21 multimedia framework is a supplement and an extension of MPEG-7. It aims to describe the "big picture" of MPEG; that is, it identifies and describes all the components and their interrelationships necessary to use multimedia resources across a wide range of networks and devices [MPEG, n.d.].

2.5 Issues Arising

Where is markup leading us to in the future? Coombs et al. [1987] pointed out that descriptive markup had been recognized as a superior markup system that was conceived as a tool for document portability, and foresaw that descriptive markup would bring improvements in the quality of digital publishing. I see Coombs' expectation is being realized through industry-wide active discussions and experiments on XML.

There is evidence that the XML community is growing and diversifying. The research study also indicates that the adoption of XML has a positive and encouraging future [Intellor, 2001]. However, the research study also pointed out a number of challenges in the adoption of XML in industry. Immature technologies and lack of qualified IT staff form the biggest two challenges, in that order. Others challenges include lack of clear guidelines and direction, lack of any overriding imperative reasons to implement, lack of understanding of XML, immature XML tools and so forth.

I discovered from MIMAS, discussed in Section 3.1.2 of this Chapter, and my research case studies discussed in Part two of this thesis that for the "old" digital libraries developed before XML existed, changing their retrospective data from SGML to XML is not a tough problem because XML is a subset of SGML. On the other hand, as far as the "new" digital libraries developed after XML are concerned, they can benefit from the outset from XML as an encoding format. The most challenging tasks for the XML-based digital libraries are that the XML family of technologies is still developing, but this situation appears to be improving.

I agree that the current state of XML-related technologies, from new initiatives to tool support, is

not keeping up with the increasingly widespread nature of XML usages. The XML Schema is one example of this, which I discussed in Section 2.3.3 of this Chapter. The major factor determining the adoption of XML and its family of specifications will depend on the quantity and quality of available tool support. Interoperability of the XML tools to create cross-platform Web applications will be an issue as well. This would need effort among XML developers. On the other hand, I have seen that XML-aware tools are being actively developed and have made noticeable progress. Institutions are more likely to use XML when the technology is becoming available with time. My three case study digital libraries give good evidence of this. The Perseus Digital Library has already adopted XML as their main technology. This is because the Perseus Digital Library is by nature a technology-oriented research testbed; yet, at the same time, they still urge staff to gain more knowledge about advanced XML technology as XML-related technologies are still under development. Michigan has made its platform compatible with XML and has scheduled the transfer of their text creation from SGML to XML. The Library of Congress is still using SGML but is experimenting with XML technology in its new projects. Although the XML environment is still maturing, there are substantial real world needs. The value of my research is therefore that it aims to provide useful guidelines to the digital library community.

Modern technology adds great value to traditional text in a number of ways. Text-based content management becomes more efficient and versatile. The display of the text and appearance could be in various formats for different purposes. Furthermore, the revision process could be finished in a much shorter time scale. The content can be updated and hyperlinks between related words are made available. As more and more texts have been moved into an electronic format, it becomes clear that XML technology is facilitating the use of content management.

The main problem in producing multimedia content would be the lack of standards and support from mainstream industry. Text-based content on the Web is going to embrace new generation techniques, that is, making use of rich markup such as XML along with metadata, while multimedia content is only just beginning to be discussed with more open standards needing to be developed and made available.

SGML is limited in terms of its capability for Web delivery and XML has been developed to overcome the limitations. This is the crucial aspect of digital library development from which digital libraries will suffer if they do not take XML into account. The XML-based infrastructure is evolving gradually. Once it is completed, the ideal Web application under an entirely open environment is not a utopian future but could be a reality. In general, it still needs a great deal of innovation and creation before XML technology becomes mature. Many people with vision have been exploring what could form an efficient and intelligent Web application such as the Web

Services. I will discuss these further in Section 2.3.1 of Chapter 8. Berners-Lee et al. [1999] thought that when XML technology had reached a mature stage, it could eventually result in the realization of the Semantic Web. I will discuss more on the idea of the Semantic Web in Section 4.1 of Chapter 5 and the potential problems of the Semantic Web in Section 2 of Chapter 10.

Chapter 3

Digital Libraries

This Chapter takes a comprehensive look at the background of the evolution of digital libraries, the collections in digital libraries and the metadata issues in digital libraries.

3.1 The Current State of Digital Libraries

The processes of scholars, students, the commercial world and the general population in finding and using information are in rapid transformation. Digital libraries have been developed to take this into account. As the use of the World Wide Web has increased, digital libraries have been built on the platform of the open World Wide Web protocol and its associated technologies. Multimedia resources which were in the earliest days of the digital library generally made available on CD-ROM because of the low bandwidth of the Internet are gaining a wider audience via the World Wide Web and are being built into applications for education, reference, and research. Digital libraries have been established by a number of sectors, both commercial and non-profit. Examples developed by the commercial sector include the digitized collections of publishers, mostly, but not exclusively, periodical literature which can be accessed through services such as INGENTA [INGENTA, n.d.] or the publishers' own Website, or the wealth of government documents available in services like the United States Government Information Locator Service (GILS) [GILS, n.d.]. All these phenomena are having a huge impact on the education and research communities.

3.1.1 Defining Digital Libraries

Academic libraries are undergoing rapid change to what might be called the hybrid library era combining print-on-paper, digital resources and other media such as film and audio. This term "hybrid library" used predominantly in the United Kingdom refers to the library which has retained traditional material but also gives access to digital library material for its users [Rusbridge, 1998]. This term is not to be confused with digital library since the digital library is very much a subset of the hybrid library. Incidentally, the origin of the term "hybrid library" is

obscure [Carr, 2001], but it was used in a JISC Circular in 1997 [JISC, 1997]. We will see one example in Section 1.3 of this Chapter of an eLib hybrid library programme.

At the same time, academic libraries are seeking to develop the networked information environment for their users. There are plenty of similarities between the traditional library and digital library in terms of information purchasing, organizing, preserving and distributing. Yet, the forms in which the information is expressed and the methods that are used to manage them make substantial differences between the two.

The term "digital library" can have various meanings for different people. Terms such as "electronic library", "networked library" and "virtual library" are often used synonymously. Originally, the term "digital library" was used by ex-Vice-President Albert Gore in the report on the role of libraries in the National Information Infrastructure (NII): "Digital Library is used here as an aggregate, implying electronic access to many sources of digital information." [Information Infrastructure Task Force, 1994]. A brief definition for digital library was given as: "a library with extensive electronic collections in a variety of forms in different locations". Another extended definition from the Association for Research Libraries (ARL) [1995] is:

- The digital library is not a single entity;
- The digital library requires technology to link the resources of many;
- The linkages between the many digital libraries and information services are transparent to the end users;
- Universal access to digital libraries and information services is a goal;
- Digital library collections are not limited to document surrogates: they extend to digital artifacts that cannot be represented or distributed in printed formats.

Notwithstanding the definitions above, the digital libraries field has broadened its scope to cover not only libraries themselves but also museums, archives, data collections and the like. The museum, archives and library professions share the same service domains on the cultural and educational landscape contributing to cultural values, learning potential, economic prosperity and social equity. In the United Kingdom, the creation of Resource: the Council for Museums, Archives and Libraries (from 2004 renamed the Museums, Libraries and Archives Council) reflects the vision that museums, archives and libraries can work together so that their invaluable contribution may be developed and sustained [Re:source, 2005]. Another case is the Arts and Humanities Data Service (AHDS), which is jointly funded by the Arts and Humanities Research Board (AHRB) and the Joint Information Systems Committee (JISC) of the British Higher Education Funding Councils. AHDS provides the higher and further education

communities with practical instruction in applying recognized standards and good practice to the creation, preservation and use of digital resources through the collection of data for management purposes, workshops, projects and series of Guides to Good Practice [AHDS, 2003].

In the United States, there are numerous institutions playing the role of data collection and broadening the scope of digital libraries in that context. Among them, the GILS is a decentralized collection of agency-based digitized documents, and the service uses network technology and international standards to direct users to relevant information resources within the Federal Government [GILS, n.d.]. In addition, the Clearinghouse for Spatial Data is a distributed discovery mechanism for digital geospatial data developed by the Federal Geographic Data Committee (FGDC). FGDC creates and manages the data elements defined in the Content Standards for Digital Geospatial Metadata, and promotes the use, sharing and dissemination of geographic data [FGDC, n.d.].

Studies are being made of the social and economic aspects of digital libraries. Technological advances bring social and economic changes. Digital libraries are social as well as technological entities. Another theory considered that digital libraries are in fact a case of an information economy due to the brokering environment [Schauble and Smeaton, 1998]. In such an information economic framework, authors, publishers and information-agents (libraries) play their roles of creating, selling and providing an information service. Libraries typically act as intermediaries between rights owners and the users. In the case of digital libraries, ease of access and costs of providing materials make the economics different from that of the traditional library. The emerging pricing and charging mechanism in digital libraries will eventually provide "transparent" services to a variety of users. The users are likely to be charged, based on the demand for different qualities of service for information access [Sairamesh et al., 1996].

Other researchers are looking at the usability of digital libraries. Much of this research has started out from disciplines other than librarianship, information science or computer systems development. In this context, the term digital library has been associated with the totality of "published" material on the Web. Some researchers remind us that when new technology is developed, words often have to be borrowed from other disciplines: this is particularly the case with new computing technology. Blandford and Buchanan [2002] draw attention to the fact that there is little understanding about what user assumptions are over what constitutes a digital library. They talk about a 'library metaphor' which in their usability studies could influence their users and enable or hinder them in their understanding of and appreciation of the features associated with digital resources.

3.1.2 The Vision and the Background

3.1.2.1 The Vision

The visions of digital libraries appeared decades ago. Bush's vision of Memex firstly inspired the digital libraries initiative: "a device in which an individual stores all his books, records and communications which is mechanized so that it may be consulted with exceeding speed and flexibility" [Bush, 1945].

Nelson [1974] described a global online library containing all of humanity's literature in hypermedia format through the vision of Docuverse, which was conceived on the foundational paradigms of what he called the WEB. The underlying hypertext paradigm and URL protocols made the WEB Docuverse technically possible. In 1965, Nelson developed a prototype, which modelled many of the concepts that make up any hypermedia system, including the WEB.

Taylor [1975] gave a hint of what an academic library would be like: "The academic library will become a true switching center, a community center in which the dynamic process of negotiating and connecting users to people, materials, and media is the heart of the enterprise... It will become a library without walls.".

Lancaster [1978] described a prosperous picture of future libraries in his prophetic book: *Toward Paperless Information Systems.* Lancaster viewed digitized information as an integral part of traditional printed information, and envisaged that network-based information services would play a key role in the 21st century library services.

The developments of digital libraries give us a vision of what a modern academic and research library would be like as described by Rowley [1998]:

- All university courseware, support materials, teaching collections are held in the digital library.
- University programmes are delivered primarily through student-centred, resource-based methods.
- The library is a member of the National Digital Library consortium and through it the Global Digital Library.
- Library staff is composed of multi-skilled learning support, guiding people comprising information, IT and academic expertise.
- Information access/study time per student is 70% electronic, 30% print.

- Library space is 70% networked study space, 30% book stock.

It is interesting to see how these visions have been realized. Digital libraries cannot provide everything that is found in a traditional library such as a place to sit and read (books or digitized materials) or the social aspects of libraries though, to some extent, aspects of the invisible college can be provided by emails, chatrooms and so forth.

Perhaps wider than that of the digital library is the influence of electronic teaching and learning which are being realized in virtual learning environments (VLEs) which include not only the published material that libraries collect but the often invisible materials such as reading lists, and photocopies of articles kept behind the counters in short loan or reference only collections. Rowley's first two points refer to these and many universities or departments within universities have yet to realize these, since they require traditionally-disposed academics to acquire new skills. Although Rowley's National Digital Library consortium is still in the future the digital library is available, a mixture of local, consortia, national and international initiatives. Library staff are often nowadays called learning services staff and may also be multi-skilled, indifferent to whether the materials they provide are traditional or digital. If, as Rowley foresaw, 70% of library materials were digitized and library space were 70% networked study space, it would mean that even more students than ever could access the materials, since they would in many cases be able to access the material from home and would not have to use 70% of the space in the library.

3.1.2.2 The Internet and the World Wide Web

The Internet is a group of networks that use the TCP/IP (Transmission Control Protocol/Internet Protocol) set of protocols to communicate. Data were originally communicated mainly by e-mail, telnet, FTP (File Transfer Protocol), and some other applications which allowed communication and file sharing across the network. The Internet has exploded in popularity on a world wide scale, with a major component of its success being the World Wide Web. The World Wide Web was made popular by an easily manipulated user interface, the web browser, in combination with a data transfer protocol the HyperText Transport Protocol (HTTP) and the HyperText Markup Language, all proposed in combination by Tim Berners-Lee to establish the World Wide Web [Naughton, 2000].

The Web has increased the world's information expectation, and that brings changes to how people use the information. The public at large are increasingly making use of online information and other forms of electronic information. In higher education and research

institutions, the process of scholarly communication is changing. Through networks and digitization, the academic community now has a wide range of networked resources available. Digital technology has expanded the availability of information to faculty members and students for research and study.

Because the majority of material on the Web is freely available, it is not necessarily associated with a library. As we saw above, the term "digital library" can refer to the totality of published material available via the Internet. But overall, traditional libraries have not been backward in embracing this new means of access to the new digitized media. For example, under the framework of the People's Network in the United Kingdom, there was starting in 2002 a major initiative to provide Internet access to public libraries across the country such that, by 2005, 20,000 terminals to the Internet had been made available in public libraries in England [MLA, 2005]. Firstly, in libraries, there have been enthusiastic efforts in library management systems over the past twenty years with the aim of providing more effective library and information services to users using non-print on paper materials, from microfilm, online information services, CD-ROMs and now the Internet. More recently, digital libraries have been seen as a next step for library management systems. For example, in 2004, a library system vendor, DYNIX, promoted a digital library to be used in tandem with its library management system [DYNIX, 2004]. The technology of digital libraries is flourishing rapidly. The Web with its rich collection of information can serve as a foundation for future generations of digital libraries.

3.1.2.3 Academic and Research Libraries and Scholarly Publishing

Chodorow and Lyman [1998] defined academic and research universities as institutions for "the production of specialized publication" with the obligation "to collect and organize the information that its faculty and students needed for their work". This obligation was historically performed by the library. Today, it could be achieved through digital libraries on the Internet. However, the material as we have suggested is not well-ordered and as such needs a gatekeeper. I think that the librarian could be the ideal person to do this because of his or her experience in the organization of information.

The Web has made publishing more accessible, and thus changed the infrastructure of information delivery to the scholarly publishing community. Libraries and publishers take access and the economic considerations into account, and are gradually moving the scholarly journal service online. At the same time, university administrators, after years of relatively static funding, have been seeking alternative approaches to the increasing cost of paper-based journals

and books, and physical space restrictions to hold them. New technology has been seen as potentially one of the best ways of alleviating the space problem and at the same time it can be used as a tool for achieving effective change in library services [Follett, 1993].

Other trends in scholarly publication include the desire to circumvent the commercial publishers; there have been numerous company mergers leaving a small number of large players in the field such as Elsevier and Blackwell. These are now able to name their prices to the academic community. The Information Access Alliance in the United States consisting of a coalition of the American Library Association and a number of other specialist library associations and a scholarly publishing association is fighting against this trend [ARL, 2003].

The academic community cannot afford constantly increasing prices. Journals are cancelled, so those still subscribing have to pay more so that publishers can assure a profit. Some journals are even introducing charges for contributors in order, so they claim, to break even [Abate, 1997; Quint, 2002]. Journals have traditionally refereed their contributions: now there are moves to provide journal articles elsewhere, in digital libraries whose content is controlled by committees which simulate peer-refereeing. The Public Library of Science [n.d.] is one of many initiatives which are providing the same service as traditional published journals and making a charge to authors, but since it is not a commercial concern, the charges are much lower. Funding from JISC has enabled researchers to publish in the open-access journals of the commercial publisher BioMed.

The e-print repository movement is another way of disseminating research whereby the academic institution hosts a repository for articles in various states of publication (draft to final commercial publication) from its own authors. A number of projects have been funded by JISC under the Focus on Access to Institutional Resources (FAIR) programme such as Theses Alive! at the University of Edinburgh [MacColl, 2002] and DAEDALUS at University of Glasgow [Nixon, 2003]. XML is one of the core technologies in the EPrints Archives at the University of Southampton [Gutheridge, 2004]. In general, there are still several hurdles to developing the approach such as copyright, getting materials, getting compliance with other systems in digital libraries and the popular support required to bring success.

3.1.2.4 Multimedia in Digital Libraries

Technological advances have made it possible to store non-text information such as photographs, maps, images, and digitized video and sound and transmit them across the Internet. These world-wide digitized resources offer a rich hypermedia working and learning environment.

These began to be used in such scientific topics as astronomy, bird migration, geography, medicine and the like. These technologies have expanded to other areas such as the delivery of culture and scholarship to classrooms, to research institutions and to individuals who are interested.

The DISCOVER project was a pilot project for the National Library of New Zealand Digital Library Programme in 2000 and it supported the school curriculum with music and the visual arts [Rollitt et al., 2002]. DISCOVER contained more than two thousand five hundred multimedia items including paintings, photographs, posters, video clips, music, essays and bibliographies, many of which reflect the Maori heritage. The primary goal of the project was to ensure interoperability and interconnection by deploying best practice initiatives such as the XML family of specifications RDF and XSLT, Dublin Core and EAD.

3.1.2.5 Teaching and Learning

Digital libraries are revolutionizing education and research methodology within academic and research communities. Global research becomes available regardless of the physical locations of the research team [Brophy and Wynne, 1997]. Distance education and instructional technologies are emerging as important new programmes for many institutions of higher education; they are a central part of the Internet2 initiative [Internet2, 2001] which seeks to provide high bandwidth to educational institutions to make possible the wider dissemination of bandwidth-hungry multimedia materials. Virtual learning environments are being introduced to enable remote teaching and learning and the core of this is the digital library or access to material in digital libraries [JISC, 2002]. Digital libraries with a wide range of networked information resources and services are ideal to be effective partners with faculty and instructional technologists in the implementation of these programmes.

MATRIX provides a good example of a project realizing teaching and learning multimedia materials in a wide bandwidth environment [MATRIX, 2003]; it is devoted to two main challenges: the digitization of sound files so that they can best be utilized by teachers, students and researchers, and the development of large-scale integrated research tools that can be developed by widely disparate repositories and freely accessed worldwide. One such example from MATRIX is "The Spoken Word: New Resources to Transform Teaching and Learning", a project which integrates the rich media resources of digital audio repositories into undergraduate courses in history, political science and cognate disciplines in the United States and Britain. It is worth noting that MATRIX Multimedia Digital Repository implements an XML database and the new metadata standard, METS schema, to manage the dynamic generation of complex,

multimedia objects.

This type of project is realizing the environment that Marchionini and Maurer [1995] were foreseeing. They considered learning in life might take place in three basic ways, namely formal learning, informal learning and professional learning. Marchionini and Crane [1994] found that such new types of learning and teaching environments provide huge potential for accelerating learning.

Traditionally, technological support and information resources for these kinds of learning have been physically separated. Digital libraries incorporate technology and information resources, allowing remote access. Teachers and students take advantage of these diverse materials, and then communicate with people outside their communities, so digital libraries facilitate integration of the different types of learning. On the other hand, this benefit is still not capable of being realized until certain technical problems are solved. The Borgman et al. [2000] research team found that the computer-based technologies in the classroom would need support not only for instructional materials but additionally more reliable support from campus network infrastructure and technical support in the classrooms. The teachers were interested in experimenting with new technology instead of traditional reliable chalk and overhead projectors, but only if the technical support was satisfied. Moreover, given the trend towards life-long learning, it is becoming essential for this kind of rich multimedia materials to be consulted from home, but few homes have sufficient bandwidth to enable it.

The learners in the classrooms may be overwhelmed by the computer-based learning methodology instead of slower-paced blackboard explanations. The question whether the digital multimedia instruction materials could really bring better learning outcomes is still an uncertain area that needs to be evaluated. Chang and Perng [2001] found at Tatung University that networked information resources were identified by research students as relatively less important information sources compared to traditional printed information sources. Of course, some factors may be the reasons such as ease of access, complexity of search interfaces and so forth.

3.1.3 Research Activities

During the last ten years, a wide variety of groups have begun to explore digital library research on a global scale. Many testbeds have been set up independently or through joint efforts. Continuing projects have aimed to consolidate results by scaling up further genuine research and practices. Here are some examples of projects involved in digital library development

research on a national or international scale.

- **NSF/DARPA/NASA Digital Libraries Initiative (DLI)** The interagency federal programme starting in 1994 was sponsored by the United States National Science Foundation (NSF), Defense Advanced Research Project Agency (DARPA) and National Aeronautic and Space Agency (NASA) [DLI, 1999]. The DLI has boosted the digital library as a worthwhile and long-term research field with emphasis not only on theoretical problems but also on practical applications. Significantly, the DLI has provided leadership in this new form of research area to advance the usability of globally distributed networked information resources through statewide and global research partnership collaborations.

DLI-phase 1 (1994-1998) invested amounts totalling over $25 million. DLI-1 focused on research areas like the handling of still images and video and the inter-working of different digital libraries. Six projects were led mostly by individuals with strong backgrounds in technical fields. However, librarians were co-investigators on several projects. The six projects were:

- University of California at Berkeley with the research topic of environmental planning and geographic information systems;
- University of California at Santa Barbara with the topic of the Alexandria project: spatially-reference map information;
- Carnegie Mellon University developed Informedia Digital Video Library;
- University of Illinois at Urbana-Champaign with the topic of federating repositories of scientific literature;
- University of Michigan with the topic of intelligent information location;
- Stanford University with the topic of interoperation mechanisms among heterogeneous services.

Among the six projects, the best documented was the Illinois Testbed at Urbana-Champaign which developed techniques for the representation and delivery of full-text engineering and physics journal articles in an Internet environment [Cole et al., 2000]. The University's Graduate School of Library and Information Science and the University Library were part of the research team. The Testbed consisted of over fifty-five thousand articles from more than forty-four scientific and technological journal titles, originally SGML-formatted and then migrated to XML. More technologies were implemented as full-text repositories have evolved and initiatives for linking and document representation have matured with time. These

technologies include many of the XML family of specifications such as XSLT, DOM and RDF. Dublin Core Qualifiers (DCQ), Digital Object Identifier (DOI) and OpenURL were used as tools for reference linking in a cross repository retrieval system. In addition, the Illinois Testbed developed several prototype data provider interfaces in the test of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and collaborated in partnership with the University of Michigan (one of my case studies) on a grant which resulted from the "Proposal to Implement a Scholarly Information Portal Using OAI Metadata Harvesting Protocols", which developed open source software tools for harvesting metadata related to cultural heritage and a search portal for the harvested data [UIUC, 2002b].

DLI-phase 2 (1998-2002) granted projects for $48 million. Phase 2 greatly expanded the support of different disciplines working in the digital libraries field. It consisted of three major components: the Research, Testbeds and Applications component; an evolving Undergraduate Emphasis component; and the International Digital Libraries Collaborative Research component [Griffin, 1999]. One of my case studies, the Perseus Digital Library, also secured funding and this grant.

Whilst a broad array of information is now Internet-accessible, human librarians are becoming increasingly sophisticated in their use of the Web. This is being replicated in the Automatic Reference Librarians for the World Wide Web at University of Washington, which was one of the DLI-phase 2 projects. It creates software agents that possess reference intelligence. The software agents know how and where to find high-quality information on the Web, acting as reference providers like human librarians do [Etzioni and Weld, 1999].

- **United Kingdom Electronic Libraries Programme (eLib)** eLib resulted from the Follett Report [Follett, 1993], which was a review of the United Kingdom Higher Education Libraries published in 1993 [Rusbridge, 1998]. eLib had an important impact overall in raising awareness of the electronic library agenda in higher education and non higher education libraries. The eLib programme characterized itself from the start as more practical-oriented than research, with a mission to provide a high quality national network infrastructure for the United Kingdom Higher Education and Research Councils communities. The eLib programme started in 1994, and was managed by a government-funded agency, the Joint Information Systems Committee (JISC). Some programmes worked in close collaboration with international agencies such as the NSF and EU (European Union). eLib Phases 1 and 2 ran in parallel and accounted for £15 million over three years to fund sixty projects in a variety of programme areas, involving over 100 higher education institutions in the United Kingdom.

Many of the issues which arose in eLib Phases 1 and 2 were taken forward in eLib Phase 3 and

their associated JISC programmes [Whitelaw and Joy, 2001]. Phase 3 ran between autumn 1997 and mid-2001, and accounted for over £4 million. The twenty projects in Phase 3 focused on three different areas which were thought would be of most benefit to academic libraries as a whole in contributing to their more economic management, regarded as necessary by Follett: these areas covered hybrid libraries, large scale resource discovery and preservation. The achievements that eLib realized were building relationships between communities such as libraries and publishers and developing the skills base.

Birmingham University Integrated Library Development and Electronic Resource (BUILDER) was developed as part of Phase 3 of the eLib programme. BUILDER developed a working model of the hybrid library which covered a wide range of activities such as staffing, technical infrastructure, authentication and authorization, electronic short loan, exam papers service, hybrid library demonstrator, digitization, copyright and licensing and so forth [BUILDER, 2001]. It is interesting to note that many of the lessons learned in BUILDER were the same as the results from the investigation in my case studies. For example, BUILDER found that managing an eLib project needed strong institutional support lack of which can create a big barrier. Otherwise, it will be difficult to fulfil properly such targets as recruiting and retaining the right kinds of staff and staff development. The project also found adopting proprietary information technology can cause problems in terms of keeping IT up-to-date. Although they did not indicate it in the report, I believe that these problems could be avoided by using the latest technologies and initiatives such as XML. In addition, in the context of eLib, the Electronic Short Loan (ESL) activity contributed to the development of the national service as a test site, and thus made possible the ESL as an on-going development pilot service as part of the Higher Education Resources ON-demand (HERON) project [BUILDER, 2001]. BUILDER finished when the project funding ceased, but HERON continues on a commercial basis and has been absorbed into INGENTA [eLibrary, 2002]. Judging by its take-up in so many institutions, it has been one of eLib's greatest successes.

Overall, it seems to me that most of the other projects from eLib were not developed to be any more than research projects and were not able to transfer from project to real world digital library practices as DLI has done.

3.1.4 Research Organizations

Fruitful research efforts have been proceeding in many independent or commercial research institutions worldwide. Most of the organizations were established to help libraries advance

through pooling resources and this is clearly becoming even more important and necessary in the digital age.

- **Research Libraries Group (RLG)** The Research Libraries Group is a not-for-profit US corporation of over 160 research-led institutions devoted to improving access to information that supports research and learning for which they use RLIN, the Research Libraries Information Network. RLG's success is based on collaboration among research repositories. RLG members get involved deeply in collective projects, and develop jointly best practices that could be used in other endeavours. A series of RLG Digital Collections projects aimed to address reformatting and access issues. RLG also launched a programme on archiving digital information primarily in developing standards, best practices and guidance for managing digital archives.
- **OCLC** OCLC, originally the Ohio Colleges Library Cooperative, which began as a cooperative cataloguing consortium, has long concentrated on efforts on issues of preservation in digital libraries, and has led the way with the metadata standards workshop series. One of its best known successes in the research area is the Dublin Core Metadata Initiative (DCMI), which has been leading the development of a set of structured metadata to support resource discovery.
- **The Coalition for Networked Information (CNI)** The Coalition for Networked Information is a US-based organization with the mission to advance the transformative promise of networked information technology for scholarly communication and the enrichment of intellectual productivity. It was founded in 1990 by the Association of Research Libraries and EDUCAUSE. CNI's meetings attract individuals not only from university libraries or computing centres, but also from publishing, network and telecommunications and government organizations. CNI's activities cover a broad range of interests, which highlight the practical application of digital libraries, collections, relationships between libraries and publishers, and policy issues of access to intellectual property. The ongoing CNI projects cover three programme themes: developing and managing networked information content; transforming organizations, professions and individuals; and building technology, standards and infrastructure.

3.1.5 Commercial Organizations

Commercial organizations have been slow to embrace digital library technology. There is an awareness that the Internet can provide profits for organizations that make use of it in an advantageous way to make good profits. However, early efforts such as NetLibrary, which provides digitized material to library users foundered because they did not optimize the algorithm for charging their users and were taken over by OCLC [NetLibrary, 2005]. Other

efforts such as eBooks.com are really digital bookshops rather than digital libraries [eBooks.com, n.d.]. Information has a cost but users do not appreciate that and have been reluctant to purchase information from the Internet, even if a way is found of charging them. All this may change. Google Inc. has taken the lead here and has agreed with four large university libraries and one public library to digitize a proportion of their collections and make them searchable online [Google Inc. 2004]. Their method of making this cost effective both for themselves and the libraries digitizing is to use this material to attract users to Google and then show them advertisements which will make money for all parties participating. It remains to be seen how successful this is.

3.2 Content in Digital Libraries

Electronic texts and multimedia resources are invaluable source materials for research and learning. Digital libraries of the future will provide access to this wealth of information media. It is worth looking at the different kinds of content which are found in a digital library.

3.2.1 Types of Content

3.2.1.1 Text

The written word has served an important function within scholarly communication that contributes to enlarging and disseminating knowledge through teaching, research and publication, and through the preservation of access to the scholarly record in libraries. By 1994, machine-readable texts had been used for research in the Humanities for over forty years [Hockey, 1994]. The mechanism of managing, storing and retrieving intellectual information ties heavily in to technological advances, from the paper-based analogue resources of the past into the digital future.

The benefits of electronic text vary in many aspects. Hockey [2000, Chapter 5] surveyed the literary uses of electronic text and thought electronic text was useful at two general levels. Firstly, it is a good tool for undertaking comparative studies of authors and text which may require large amounts of time to analyze manually. Secondly, when testing hypotheses, it provides concrete evidence to support or refute hypotheses or interpretations which previously had no more validity than mere hunches. Hockey [2000, Chapter 6] also looked at its uses in linguistics which fall into two categories: in the study of linguistics, analysis of large corpora to

determine how a language in general (in a particular place or time, for example) is used; and in the study of language by language learners (usually at a high level), by interpreters and translators to determine how a particular word is used, since searches can be made for a particular word in a large volume of text. All these benefits were found from using a digital library.

There are a number of different ways for migrating materials from analogue formats to digital formats. The simplest is to convert the document to an image file. Secondly, it is possible to convert the data to ASCII or Unicode, which means it can be incorporated into other documents, as if the document had been produced on a word processor in the first place. ASCII and Unicode, however, do not incorporate any standard for formatting apart from line break characters. There are a number of ways of doing this, and markup is probably the most successful. Marked-up ASCII text can potentially represent any feature of a published document such as indentation, emboldening, italicization, pagination and so on. So, the use of a flexible markup scheme to create exchangeable and structural electronic data goes some way towards solving the problem of formatting. But, it also goes beyond formatting and adds value to the text because it can make it possible to search through the text in a way that is not possible with a pure image file. However, because of the flexibility of XML, it is possible, using XML, to add links to the text which could, for example, link a word in the text to its definition in a dictionary, and add even greater value through XML. There are other uses. Smith [2001] at the Perseus Digital Library described how XML made it possible to make public the kind of data for which manuscript annotations have been used in the past (known as Grangerization after James Granger who published a work with blank pages for readers' notes). Thus, XML can enable linking which cannot be achieved by traditional paper, automatic distributed mono-directional linking and gathering relevant information as to what the reader is reading such as citations and annotations. Muller and Beddow [2002] also stated that XML provides the possibility for a search in one dictionary to be passed to another XML-compliant dictionary to fulfil the search.

Full text journals are another medium for which digitization is bringing advantages. Many libraries do not allow journals to be loaned: digitization and access to the digitized documents from outside widens the access possibilities as well as enabling more readers to read a journal issue at the same time. Moreover, in universities with many different libraries, it can save on the purchase of multiple issues, something which is causing the publishers concern. On the other hand, now that journals can be seen in digital libraries, publishers are making deals with groups of organizations often in countries outside Europe and North America, that hitherto could not afford each to buy copies, allowing them access to large ranges of material which were too expensive to send to them when transportation as well as printing costs were taken into account,

as was the case with printed materials. For example, the CONsortium on Core Electronic Resources in Taiwan (CONCERT), which takes advantage of the growing popularity of Web-based full-text documents, does its best to fully exploit the economies of group purchase and resource-sharing [National Science Council, 2004].

3.2.1.2 Image

Given the increasing creation and dissemination of data in various formats on the Web, one cannot neglect image material as a useful electronic resource contributing largely to the presentation and visual impact of educational packages. As the *AHDS Guides to Good Practice* state, a digital image is one form of image, simply another form of representation. There are good practices in producing digital images in the same way as there are good practices in painting, photography or sculpture [AHDS, 2000]. Scanning, storage and representative technologies are advancing rapidly, and the choice of approach has been affected by several possible factors: the state of technology at the time of capture, characteristics of the original document, physical characteristics of the original material or surrogate which could be scanned such as microfilm, and the cost.

A number of efforts worldwide have investigated the creation and use of digital resources including the UK Arts and Humanities Data Service (AHDS). The *AHDS Guides to Good Practice* is a series of guides which indicate best practice and standards in digitization in the areas of the Arts and Humanities [AHDS, 2003]. The guides aim to work through all aspects of the digital processing cycle starting with the creation of the material and providing assistance in planning, production, delivery, use and preservation of all forms of high quality digital materials. And this is complemented by organized programmes which include publications, workshops and consultation.

Compared to the large number of sources and systems for electronic texts, there are fewer digital libraries devoted to digital pictures. Many digital libraries include images in their documents. Some textual material is digitized as images such as manuscripts because it is important for researchers to see the original, or something as near to the original as possible. JSTOR stores both image files and text files side by side, so that the end user can search the text files using normal text searching mechanisms and retrieve the image, so that the page looks identical to the original printed journal page and would, of course, retain any original images that were embedded in the text [Holden, 1998].

A panel at ACH/ALLC'99 International Humanities Computing Conference recognized recent

technical developments in mechanisms of delivery and browsing image data, and thought that there was the need to treat image data as structured data rather than simple supplements to machine-readable text, so that digital images can be regarded as having the same degree of intellectual scrutiny as text [Drucker et al., 1999]. Meanwhile, a consensus was reached at CNI/OCLC Image Metadata Workshop that the fixed/static/bounded image was characterized as a fixed document-like object, similar to the text-based document-like object; therefore, a resource description model was to be developed to support network-based discovery [Weibel and Miller, 1997]. Other data formats that were thought to be treated in the same way included movies, musical performances, speeches, and other information objects. I will discuss this in the following section.

In September 1999, the California Digital Library (CDL) [CDL, 1999] developed a set of digital object standards in metadata, content and encoding, and a set of standards for digital image format. These documents have been reviewed and updated annually. In the documents, metadata required for simple digital objects such as a single archival image, and complex digital objects such as a book, were outlined and explained. However, metadata systems such as TEI Header cannot be used alone for digital object encoding as this is not as straightforward as electronic texts. I will discuss this further along with my three case studies in Section 2.1.1.2 of Chapter 8.

3.2.1.3 Dynamic and Complex Objects

Many of the digital objects that are now being considered for digital library collections are non-document-like objects; they cannot be represented as static files of data [Arms, W.Y., 2000] such as CAD/CAM, geographic information, databases and computer programs. For example, data from a scientific sensor or an image from a video game will have a different content every time it is presented to the user. Other library objects which the digital library regards as single library objects can be a combination of different types of object, in which there may be complex relationships one to another. These distinctions are often blurred. For example, an article in an online periodical may be stored on a computer system as several files containing text and images with complex links among them.

Multimedia content has some characteristics that are fundamentally different from text. Therefore, the models and tools that are developed for text cannot be readily applied to multimedia. The relationships between these components are multifaceted including temporal, spatial, structural and semantic. Any descriptions of a multimedia resource must account for these relationships. To manage well such complex entities is a considerable challenge. Ossenbruggen thought managing such complex entities, richly annotated multimedia

presentations, would be a key pre-requisite for the multimedia variant of the Semantic Web [Ossenbruggen et al., 2001]. For example, time-based media such as audio and video would need "mechanical" metadata for controlling processes such as synchronization, which would need to be linked to higher-level descriptive metadata through abstractions.

3.2.2 Long-Term Preservation and Reuse

There has been a rising level of interest throughout the library and archive communities in the use of digital imaging for preservation reformatting as evidenced by many conferences devoted to this topic during recent years. The digital collection must guarantee the longevity and authenticity of information that facilitates reuse and enhancement by a broad scholarly community in a networked environment. Therefore, preservation, management and access to these materials over time have been acknowledged worldwide as a substantial challenge in digital library research [Day and Powell, n.d.].

The underlying issues of preservation of digital objects were first elaborated by a commission of the Council on Library and Information Resources (CLIR), the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG) Task Force on Archiving of Digital Information. They worked together to identify preservation policy framework and research agendas, trying to seek more reliable and cost-effective methods for digital preservation. They felt that the challenge for digitized archives was to retain intellectual content so the ideas available in the end were identical to those in the original object [Garrett and Waters, 1996]. I learnt from my research interviews that this will be the methodology in the National Digital Information and Infrastructure Preservation Program (NDIIPP), which the Library of Congress was planning to conduct. NDIIPP is a consortium effort that brings together institutions, commercial technology companies and creative communities to develop solutions to the preservation of digital content.

In the United Kingdom, the JISC and the British Library firstly addressed the issue of the preservation of digital media by holding a national conference in Warwick in November 1995, where a number of action points were identified [Fresco, 1996]: more work was needed to specify the preservation process; research was needed to establish benchmarks for determining when a digital document was deteriorating to the extent that action is needed for its preservation; and attempts should be made to increase the pool of digital skills. In Washington DC, the National Research Council's (NRC) Computer Science and Telecommunications Board (CSTB) strongly urged the Library of Congress to take a leadership role in the development of digital preservation technologies and in the creation of relevant metadata standards and practices

[National Academy of Sciences, 2000, pp. 105-121].

The National Information Standards Organization (NISO), the Council on Library and Information Resources (CLIR), and the Research Libraries Group (RLG) sponsored a workshop to examine technical information needed to manage and use digital images that can reproduce a variety of pictures, documents and artifacts [NISO, 1999]. A metadata set was proposed by an experimental project of the Library of Congress and the Corporation for National Research Initiatives (CNRI) in association with the earlier workshop.

Digital documents are vulnerable to loss probably to a greater extent than printed materials via the decay and obsolescence of the technologies. To avoid "technological quicksand", effective preservation technology solutions need urgently to be found [Rothenberg, 1999]. There has been an increasing awareness that metadata would play a key role in digital preservation, whether it was technology preservation, emulation or migration [Granger, 1999]. The preservation metadata should include data about file formats, software and hardware platforms, and can also record information about authenticity and rights management issues [Day, 1998a].

Several initiatives have attempted to identify preservation metadata elements. The Research Libraries Group's Working Group on the Preservation Issues of Metadata assessed the preservation and metadata requirements of digital imaging technology by examining two metadata formats, the Dublin Core and US MARC-based core record standard, so that the RLG could specify the extra metadata elements needed that would be important to serve preservation needs [RLG, 1998].

The efforts of digital preservation have been explored by numerous projects. The CEDARS (CURL Examples in Digital ARchiveS), a project funded by the British eLib programme and managed by the Consortium of University Research Libraries (CURL), aimed to address strategic, methodological and practical issues and provide guidance for libraries in best practice for digital preservation [CEDARS, n.d.]. The major component of the work encompassed by the CEDARS project was the development of a metadata framework which would enable the long-term preservation of digital resources. An XML DTD was developed to express the metadata elements within its Archive Information Package (AIP).

Significantly, the CEDARS project adopted the Open Archival Information Systems (OAIS) reference model developed through ISO and the Consultative Committee for Space Data Systems (CCSDS). The OAIS method was endorsed strongly as a metadata framework to support the preservation of digital objects by the OCLC/RLG Working Group on Preservation

Metadata in June 2002 [OCLC, 2002]. OAIS provides a comprehensive framework and vocabularies for establishing a complete archival system [CEDARS, n.d.]. From the research interview, I learnt that the combination of new metadata systems METS and OAIS was expected to be one of the core technologies in digital preservation. The CEDARS Access Issues Working Group produced a preliminary study of preservation metadata and the associated issues surrounding it, in which they recognized the importance of metadata creation and maintenance in ensuring the continuing of digital information objects [Day, 1998b].

3.2.3 Digitization as a Means of Preservation

Digitization is a means of preservation in its own right. Many efforts have been made to evaluate digitization as a means of preserving, for example fragile materials, by means of digital copies which will be consulted in place of the original. Five preservation methods associated to different scenarios (such as when to use content migration, or when to use digital archeology) were developed at the spring 1998 meeting of the Coalition for Networked Information (CNI) during a panel discussion on "Digital Preservation." [CNI, 1998]. The methods had proved useful in applying the conceptual framework of the "repertoire of preservation methods" adopted by the American Memory project at the Library of Congress [Arms, C.R., 2000]. The problems of preservation of the digital objects produced for the purpose of preservation are the same as for those materials which are born digital which I discussed in the previous section.

3.3 Metadata and Digital Libraries

Metadata is structured information that describes, explains, and locates information resources. The definitions of metadata will be discussed more fully in Section 1 of Chapter 5.

Carefully designed metadata is of great value in the digitized library in both the short and long-term. The value of future digitization and distributed information initiatives will crucially depend on the creation and management of the metadata. Gilliland-Swetland [2000] pointed out some key issues that must be resolved when developing digital information systems and objects:

- identifying which metadata schema or schemas should be applied in order to best meet the needs of the information creator, repository and users;
- deciding which aspects of metadata are essential for what they wish to achieve, and how granular they need each type of metadata to be, in other words, how much is

enough and how much is too much. There will probably always be important tradeoffs between the costs of developing and managing metadata to meet current needs, and creating sufficient metadata that can be capitalized upon for future, often unanticipated uses;

- ensuring that the metadata schemas being applied are the most current versions.

The metadata issue can become complex as any modifications, copying or reproduction of the original digital objects can change the associated and inter-related metadata, and that can bring difficulties to the management of the materials.

There are also various levels of granularity of metadata, for example, a cataloguing record of a digitized book in a digital collection; a collection-level description in an online directory of a digital library; a description of a virtual collection created by cross-repository searching.

It is important, therefore, to deploy a recognized and appropriate metadata set wherever possible to ensure interoperability. Not all the metadata schemas include all the different types of metadata, so it may be necessary to combine more than one metadata schema as a particular metadata application, or to extend an existing metadata schema with local elements. For example, the XML-based Linking Encoded Archival Description (EAD) to Electronically Retrievable Sources (LEADERS) project in the School of Library, Archive and Information Studies at University College London has developed a structure for the LEADERS Schema, which is based on EAD 2002 with item level descriptions but has incorporated elements from the TEI and NISO Metadata for Images in XML (MIX) elements within EAD's <altformavail> element referenced through XML namespaces. This approach allows both the original archive document and the digital forms of the archive document to be adequately described [Sexton, 2003]. In addition, the XML-based Encoded Archival Context (EAC) is also utilized for authority records.

3.4 The Challenges of Digital Libraries

Digital libraries have been developed for over a decade. Some of the visions of digital libraries which experts described in the past have been realized because of technological progress. But we still do not know to what extent digital libraries will develop in the future, or what kinds of services digital libraries will provide. Perhaps librarians working in digital libraries and computing professionals have different views on this because of their different backgrounds.

Digital libraries are revolutionizing education and research methodology within academic and research communities. The utilization, management and control of electronic quality information content have become major issues for 21st century scholars and researchers. Easily accessible, digitally formatted information has become possible. What are the implications for academia? And how can higher education adapt to distributed-learning? I think that these are questions which people in senior positions need to consider seriously in order to plan ahead for the future of the electronic educational network.

How should librarians manage the electronic content implications for libraries? How could they serve their diverse patrons and fulfil their mission as repositories of organized knowledge? Until the standardized format and associated issues relating to electronic publishing have been resolved, librarians need to keep aware of new technology developments, in particular XML. Librarians need to take into account the expectations of and the acceptance from users of the electronic content that could possibly affect the concerns I discussed earlier, and at the same time communicate and recommend their observations to the electronic publishing industry as well as library users.

When the staff of the Library of Congress identified ten challenges in building the digital library of the 21st century [Library of Congress, 1998b], they saw that metadata was the key to resolving these challenges. As Anderson [1999] at the Library of Congress pointed out, interoperability and metadata are the main components for building global networked digital libraries. Metadata is important because it is used to provide access globally, and to find and describe the digital content. It would be also important that metadata is encoded in a universal, globally accessible format which can potentially promise longevity, flexibility, compatibility and interoperability.

In a technology-focused digital library development environment, perhaps we need to rethink the role of digital librarians. What is the true value of the digital libraries? Technology is merely a tool to achieve the digital library vision. I feel it should be the digital librarian (a librarian who fully understands the digital environment) and not the computing professional who is the key to joining together technology and traditional library services, and who would be able to enlarge the boundaries of library services into a global perspective of the digital library village. If so, then more investment in digital librarians should be seriously considered. In Chapter 9, I discuss this aspect further when I discuss developments in digital library administration.

Chapter 4

Storing and Managing Structured Documents

Just as a traditional library cannot function efficiently without a catalogue, the core of the digital library is its mechanism for querying and retrieving large volumes of structured and not-so-structured data. Digital libraries use database management systems extensively for the purpose of managing and querying large volumes of structured data. The seemingly inexhaustible march of the Web revolution has exposed more and more developers to database issues because of the desire for ever more dynamic Websites. XML has had the effect of increasing awareness of data design in the commercial world and in the field of digital libraries as well. This is probably due to the fact that XML data are human readable and similar to normal text but with the addition of XML tags. People involved in contributing to XML-based systems become aware of the way the data are structured, whereas those contributing data to database management systems are unaware of the way that the data are held in the Relational Database Management System (RDBMS).

Most digital libraries separate their storage and retrieval data; the retrieval data are structured data with tagged fields; the data are stored as less structured data such as the data within paragraphs of authors' text. If XML or indeed SGML is used, the data which are required for retrieval (that is, author, title) can be extracted from the XML or SGML by means of the tags. This illustrates one of the benefits of XML: data in an ordinary text document can be tagged with XML and they become more structured, thus increasing the potential for retrieval of the text within the document.

With XML, developers not only can define tags to represent different database fields from specific domains, but also the data format is not bound to particular script languages, authoring tools, or delivery engines but is presented in standardized, vendor-independent ways. Furthermore, XML is database-neutral; there is less of a technical problem in connecting databases with heterogeneous data sources. The leading database vendors have targeted building XML-based applications such as XML-enabled Web server products. Their common strategy is to exploit XML technology to productively build and cost-effectively deploy reliable and scalable Internet applications [Muench, n.d.]. However, before these products are adopted by a mass market, we still need time for a more robust XML infrastructure to be accomplished.

This Chapter describes the concept of semi-structured data, XML technologies including XML data model and XML query languages when they are implemented in database management systems and XML databases. Two mapping technologies between databases and XML structures are also discussed. In the remaining parts of this Chapter, I demonstrate a three-tier architecture using an XML-aware relational database to illustrate a number of methods on data manipulation for presentation processing. The advantages of the XML solution in association with relational databases are discussed.

4.1 Structured Documents; Semi-Structured Data

The growth of the Internet and the emergence of XML have motivated research in the area of data models, query languages and systems for semi-structured data [McHugh et al., 1997; Fernández et al., 1997; Penn Database Research Group, 2001]. Abiteboul [1997] defined semi-structured data as data that are neither raw data, nor very strictly typed as in conventional database systems. However, this definition is approximate. In XML, the document structure is formed via the element-and-attribute tree with recursive hierarchies and repetition, while in the relational database world, the structured XML document is regarded as semi-structured data because the structure is not as rigid, regular, or complete as the table structure required by traditional database management systems.

4.2 Storing XML Data Model

4.2.1 Representing Relational Databases

Relational databases store data broken into multiple tables, each of which stores related data. Keys are used to create references from one table to another. Using the join mechanism, multiple tables can be joined, so that a single set of output is returned. The limitation of storing XML data in relational databases is that the basic characteristics of tables in relational databases and hierarchies in XML are different [DeRose et al., 1998]. To store XML data in a relational database, one needs to break the XML elements into the tabular architecture. But this is not well represented in the hierarchical, interconnected nature of XML content, as the structure and the semantics of an XML document are very likely to be lost particularly in the types of documents which are found in digital libraries. Although relational database management systems provide

powerful features like *key* and *join*, they scale poorly in large data volumes, concurrent editing, or a link management and navigation environment, since these require complex and large numbers of tables and joins [Hogan, 1997].

4.2.2 Representing Object Databases

By contrast, the structure of an XML document in many ways mirrors the structure of an object database and thus performs better in these areas. The object-oriented data model is identity-based. An object can be referenced via the Object IDentifier (OID). Relationships among objects are established using pointers. The instances of the database corresponding to a schema consist of a collection of trees whose nodes are records; each tree is called a database record [Ullman, 1988]. Unlike the relational data model, the object-oriented data model allows the storage of all information concerning an entity as a single database object, which leads to a direct correspondence between a real world object and its database representation.

4.3 Database Approach: Relational versus Object

4.3.1 Database Background

The relational database model was created by E.F. Codd in 1969 at IBM. The relational database system is good for processing large amounts of simple data and data retrieval, yet provides little support for data manipulation [Getz et al., 1994]. The object database revolutionized database management development in the 1980s. The object database system is good at handling complex relationships among objects and data manipulation such as a large number of many-to-many relationships, but provides little support for data persistence and retrieval [Cattell, 1992]. Relational database systems are more efficient when working with fixed length fields within fixed length records. Catalogue data and full-text data do not conform to this model at all. Catalogue data have many fields which may be repeatable and are of variable length [Gredley and Hopkinson, 1990, pp.44-52]. Full-text data by their nature are unstructured and difficult to convert to fixed length fields. However, as time has gone on, the RDBMS model has become more complex and is gradually becoming better able to cope with catalogue data, larger quantities of text or binary large object (BLOB) [Sun Microsystems, 2001].

4.3.2 Comparison of DBMS Architectures

Figure 4 illustrates the comparison of the two database management systems. Object database management system (ODBMS) provides an architecture that is significantly different from the relational database management system (RDBMS). An object database management system maps transparently with the application programming language, while a relational database management system needs explicitly to copy and translate data between database and programming language representations. An object database management system makes database objects appear as programming language objects, and does not need a separate data model language.

With the increasing popularity of the multiformat Web data available on the Internet, developers have re-evaluated the functional features of object database management systems because they provide a more efficient way of tackling multimedia information than relational database management.

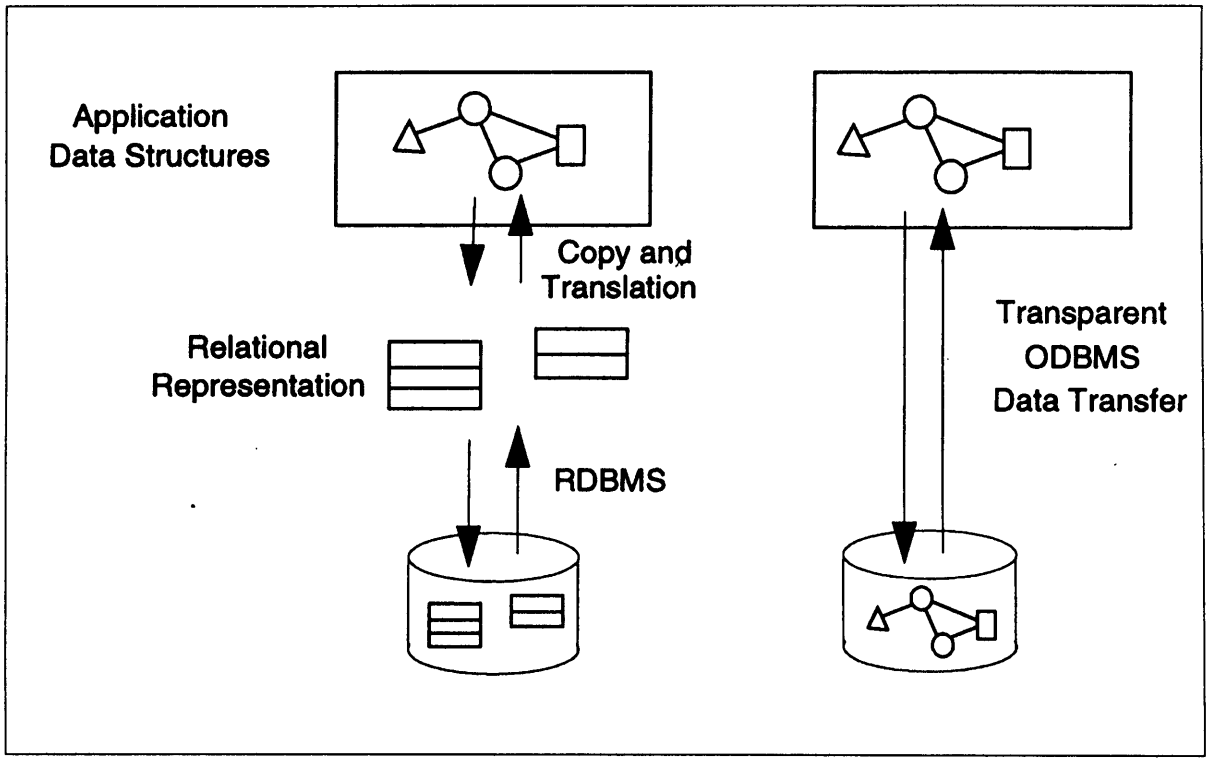

Figure 4: Comparison of DBMS architectures [Cattell, 1996]

4.3.3 SQL Extension – SQL3 (SQL99)

Relational database systems have proven the most widely adopted database systems in the databases market [Piattini and Diaz, 2000]. Much of the success came from standardization (in this context also called normalization) started with database modelling and the implemented Structured Query Language (SQL) and were endorsed by the International Organization for Standardization (ISO) and the American National Standards Institute (ANSI) [Cattell, 1996]. The high acceptance of the SQL standard (ISO/IEC 9075:1992) led to the strong support of portability and interoperability between systems that promised the achievement of the relational approach. Moreover, to support object modelling, a new extension of SQL – SQL3 (also known as SQL99) has been working in which it addresses the requirement for objects and 'object identifiers' in SQL, and also specifies supporting features such as encapsulation, subtypes, inheritance and polymorphism found in object data models [Manola, 1997; Cattell and Barry, 2000].

The two database systems are likely to have complementary strengths; however, relational vendors realized that the concept of object was important. A new extended version of relational database technology has emerged as the "object relational" database management system (ORDBMS) [Lozano, n.d.; Silberschatz et al., 1997], providing a single environment for traditional business transactions, multimedia data and complex structures. In the following sections, I will discuss how the mapping technology is applied in the "hybrid" object relational database management system, and the XML aware databases that are well known to the developers.

4.4 Mapping Between Databases and XML Structures

Much of the data already stored within relational databases have an important role in providing content for Web-based applications, making a relational database to be XML-aware; hence, the mapping issue between relational databases and XML structures is becoming one of the research topics for the XML community [W3C XML Core Working Group, 1997]. Two most commonly known mappings are table-based mapping and object-relational mapping [Bourret, 2001]. Both mappings are round-tripping; that is, they can be used to transfer data both from XML documents to the databases and from the databases to XML documents. I take an imaginary example from the real world of course books taken out of the University College

London Library and put into a Teaching Collection (or Reserve Collection) for students on that particular course.

4.4.1 Table-Based Mapping

Table-based mapping views the document as a single table or a set of tables, and the document structures would be either of these:

Table teachingCollection

course_code	course_name	lecturer	time
P037	XML	Susan Hockey	Thursday, 10-13

```
<teachingCollection>
<course>
<course_code>P037</course_code>
<course_name>XML</course_name>
<lecturer>Susan Hockey</lecturer>
<time>Thursday, 10-13</time>
</course>
.....
.....
```

OR

```
<teachingCollection>
<courses>
<course course_code="P037"
course_name="XML"
lecturer="Susan Hockey"
time="Thursday, 10-13"/>
.....
.....
```

Note the column data can be represented either as text-only elements or attributes. Comparing the two approaches, the attribute approach has the advantages of smaller document size and less

programming complexity, while the element approach has the advantage of ordering. Attributes are unordered. The element order sometimes may have meaning for documents; on the other hand, the ordering of elements comes with the penalty of programming complexity [Williams et al., 2000].

Table-based mapping is simple; it is easy to write code based on it, and is useful for simple document mappings. However, it gives no support for complex data structures such as entity reference, datatypes information and so forth.

4.4.2 Object-Relational Mapping

Object-relational mapping allows more complicated mapping; it models an XML document as tree objects that are specific to the data in the document, and then maps these objects to the databases. Consider the following XML document:

```
<teachingCollection>
 <course course_code="P037" course_name="XML" lecturer="Susan Hockey"
       time="Thursday, 10-13">
     <book course_code="P037" title="XML for the World Wide Web: visual
          quickstart guide" au_lname="Castro" au_fname="Elizabeth"/>
     <book course_code="P037" title="The XML Handbook"
          au_lname="Goldfarb" au_fname="Charles F."/>
 </course>
 .....
```

First, I map the document to the objects:

```
Object teachingCollection {
      course_code ="P037";
      course_name ="XML";
      lecturer ="Susan Hockey";
      time ="Thursday, 10-13";
      book = {ptrs to Book objects};
}
      object Book {
      course_code ="P037";
      title ="XML for the World Wide Web: visual quickstart guide";
```

```
        au_lname ="Castro";
        au_fname="Elizabeth";
}
        object Book {
        course_code ="P037";
        title="The XML Handbook";
        au_lname ="Goldfarb";
        au_fname="Charles F.";
}
```

And then I map the rows in the following tables:

Table teachingCollection

course_code	course_name	lecturer	time
P037	XML	Susan Hockey	Thursday, 10-13
.....			

Table Book

course_code	title	au_lname	au_fname
P037	XML for the World Wide Web: visual quickstart guide	Castro	Elizabeth
P037	The XML Handbook	Goldfarb	Charles F.
.....			

For DTDs and W3C Schemas, a complete mapping to object schemas and then to database schemas is available [Bourret, 2001]. A wide range of middleware tools, database management systems, or XML-enabled servers adopt the concept of object-relational mapping providing XML capabilities, products such as Microsoft SQL Server XML Features, Oracle 8i and higher and so on. I will discuss this further in Section 6.2 of this Chapter. Additionally, various database vendors and the open source community have released object-relational mapping tools for those early adopters that want to take advantage of transparent persistence but wish to remain with a relational database [Rhyno, 2002b]. Using object-relational mapping would improve performance over using an embedded SQL [Barry & Associates, n.d.]. Rhyno [2002a] discussed an open source XML object-relational database mapping tool, Castor, in building his database and experienced a smooth implementation. However, these products implement

different levels of object-relational mapping technique, and none of them is guaranteed to provide one hundred per cent fidelity when the technique is put into effect. The problem of this is that the information may be lost during the conversion process.

4.5 Query Language for XML

Semi-structured data is becoming more and more common. A frequent scenario for this is the data found on the Web; data on the Web usually come from several heterogeneous sources from loosely structured documents like HTML pages to very structured information like relational databases, or others somewhere in between. As more and more of the critical data are stored in XML documents, collections of XML files will be eventually accessed like databases; hence, the ability to query XML documents becomes an essential aspect of effectively utilizing XML documents.

The research from DeRose [1999a] and McHugh and Widom [1999] noted that XML documents have a combination of the properties of "structured" database data and "unstructured" natural language. This is because XML documents are structured, but on the other hand, the data buried in those XML elements is in deep hierarchies and unpredictable orderings and repetitions. This implies that developing a query language for XML documents could be potentially a difficult task. This can be seen in the following sections.

There has been a great deal of activity in recent years in proposing new semi-structured data models and query languages for this purpose [Abiteboul et al., 1997; Deutsch et al., 1999; Robie, 1999]. A member of the XML family of specifications, XPath, is the XML technology most frequently used for querying XML documents, and it can be loosely considered as an equivalent to SQL in a relational database.

The eXtensible Stylesheet Language Transformations (XSLT) were originally designed to allow users to write transformations from XML to HTML, thus making it possible to present an XML document via a normal Web browser. To achieve this end, the XSLT uses XPath to select parts of an XML document, where XPath is a query language used to navigate nodes of a document tree. As a query language, the tree data model of XSLT accurately corresponds to XML's, however, there are disadvantages when XSLT works as a query language in a huge database environment, for example, developers have to lock entire documents before the XSLT can be executed which is inefficient [Pawson, 2001].

XML-Query Language (XML-QL) is a query language for XML. It was developed by researchers from academia and industry, and has been sent as a submission to the World Wide Web Consortium. XML-QL has remained as a NOTE status made available by the W3 Consortium for discussion only since its first submission on 19 August 1998. XML-QL was closely inspired by both traditional query languages, such as SQL, and by existing research of query language for semi-structured data. Bonifati and Ceri [2000] found that XML-QL was able to play the same role as high-level SQL standards and languages in the relational world. Although the development of XML-QL did not move forward with the rise of XML, it laid a good base that was applied by the later developed W3C XML Query language (XQuery), which is discussed in the section below.

4.5.1 XML Query (XQuery)

XQuery is a W3C XML query language that is designed for processing XML data (including databases) whose structure is similar to XML. XQuery defines a number of different types of expressions which are the key to XQuery. Every query contains one or more query expressions [Malhotra et al., 2003]. I herewith give some examples using the previous XML file whose root element is <course>. The following example is creating a simple Web page that just lists the titles using "let" and "for" expressions to define variable definitions:

```
<html>{
   let $course := document("213.253.17.123/essay2.xml")/course
   for $bk in $course/book
   return <h2>{$bk/title}</h2>
 }</html>
```

Note the relationship with XPath; that is, all XPath expressions are also XQuery expressions.

Both XSLT and XQuery use the XML Path Language. XSLT is useful for expressing simple transformations, but XQuery could often be more compact with more complicated stylesheets, especially when dealing with programming.

The next example shows that XQuery has the ability of programming languages like Java, C++ and other languages. This invokes the process-title function if the value of $ECP is an element whose tag name is title.

```
if ($ECP instance of element title)
 then process-title($ECP)
 else ( ) {--nothing--}
```

XQuery is developed to support the work in the XML family [Chamberlin et al., 2003]. For instance, XQuery is namespace aware and has schema availability although, as of 2005, not all XML Schema components are yet fully supported. In an XML-based digital library environment, XML represents all types of information formats. XML needs a universal query language which should have the ability to describe disparate data sources. The query language techniques should focus on extracting data from large XML documents, for translating XML data between different ontologies (schemas/DTDs), for integrating XML data from multiple XML sources, and for transporting large amounts of XML data to clients or for sending queries to XML sources.

Unfortunately, the query language evolution has not quite kept up with the increasing popularity of XML discussed in Chapter 2. As Widom [1999] pointed out, the true requirements for XML query languages would not be known until a significant number of data-intensive XML applications are built. Yet, XQuery language has the potential to be widely-supported in database management systems (DBMS) and other software products, as it is XML-aware and has the required features for a query language.

4.6 XML Databases

This section is intended to give an overview of the types of existing database management systems in the market that are able to store, manage and retrieve XML data. I roughly categorize them into two main approaches: native XML databases and XML-enabled databases. I discuss the differences in general between the two with practical database management systems as an example.

4.6.1 Native XML Databases

Native XML databases are databases designed especially to store XML documents, and their internal model is based on XML [Bourret, 2004]. Native XML databases, also known as hierarchical databases or semi-structured databases, provide an XML engine that is able to store and retrieve XML data in their native format, and thus could offer better performance than other

XML-aware database systems. Firstly, the native XML database supports XML-specific capabilities, such as XML query languages, and would usually be faster while retrieving data. Developers also gain a lot of flexibility through the semi-structured nature of XML [Apache Software Foundation, 2005]. The benefits of such are the ability to store semi-structured data, and accepting, storing and understanding any XML document without prior configuration, that is, documents which contain no DTDs or schema (well-formed documents). And this offers a substantial advantage in applications such as Web search engines, where no single DTD or set of DTDs applies to all the documents. Normally, when transferring the data in an XML document to a relational or object-oriented database, the system would require first the creation of a mapping and a database schema, thus slowing down the performance. This needs to be done when extracting metadata from XML when building the index to the documents in a digital library. In addition, a native XML database does not require any particular underlying physical storage model; that is, it can be built on a relational, hierarchical, or object-oriented database, or use a proprietary storage format such as indexed, compressed files [Staken, 2001].

As of 2005, the majority of the native XML databases on the market are commercial proprietary database types, though few of them are open source such as those developed as research projects. For instance, Apache Xindice is the continuation of Apache Software Foundation project that used to be called the dbXML Core [Apache Software Foundation, 2005]. It is a semi-structured native XML database written in Java. The Medlane project at Stanford University Medical Center has experience with several native databases products for storing and retrieving XML bibliographic records, and has found Xindice promising [Clarke, 2001].

4.6.2 XML-Enabled Databases

XML-enabled databases are databases (usually relational) that contain extensions for transferring data between XML documents and themselves [Bourret, 2004]. They put an XML mapping layer on top of their relational engine to be able to store and retrieve XML data. Unlike native XML databases, they can store XML documents without knowing their schema (DTD). XML-enabled databases could generate schemas on the fly, although this is impractical in practice, especially when dealing with schema-less documents. In addition, data retrieved as XML are not guaranteed to have originated in XML form. Data manipulation may occur via either XML specific technologies such as XPath, or other database technologies such as SQL. The fundamental unit of storage in an XML-enabled database is implementation dependent. The XML solutions from Microsoft and Oracle as well as many third party tools fall into this category.

Microsoft SQL Server is an XML-aware enterprise relational database management system; it creates an XML document using a series of SELECT statements. The mapping schemas specify an object-relational mapping between the XML document and the database, and support sophisticated queries using a subset of the XPath language. In the next section, I experimented with the strengths of SQL Server 2000 through the Ching example.

Oracle is another relational database system used very much in the commercial world. Its version 8i and later versions support XML-enabled object views over relational data. The Oracle 9i database is XML-enabled to store, search and retrieve XML natively. With this new technology, users can use the new datatype called XMLType. XMLType stores XML documents as Character Large OBjects (CLOBs), which is a new approach for traditional relational databases. The New York Digital Library Team found these new functionalities in Oracle 9i helpful to improve the workflow but learnt from practice that Oracle could only work agreeably with certain open source software programs [Myrick, 2002].

4.7 Ching Digital Image Library Project

In a digital library environment, with the relational databases combining the XML/XSL approach, much presentation processing on the server side has been shifted to client side. The developer can manipulate the source XML data on the client, and save them back to the server in their native format. In this section, I present my experiment: the Ching Digital Image Library, which combined HTML and XML technologies discussed in Chapter 2, an XML-enabled database discussed in Section 6.2 of this Chapter and digital imaging techniques which will be discussed in Section 1.1.3 of Chapter 8.

Background and technical requirements

The Ching Digital Image Library was an online collection of high-quality digital images of works of art. The time period represented was a range from around AD 1736-1795, in the Ching dynasty, in China. This collection portrayed the workmanship of the Ching dynasty with the intention of providing an avenue for the understanding of the Ching period costume accessory system. The Ching project was conducted over half a year with dozens of experiments with different methods. That many different methods are valid is one of the main messages I would like to deliver through this project.

The images were scanned using an EPSON colour scanner and its accompanying software. The

files were saved in TIFF format for later conversion into JPG images for Web display. Each image was given a unique ItemNo (ID) consisting of two letter and three numbers. For example, BL237 stands for item 237, and was categorized to the subject BRACELETS.

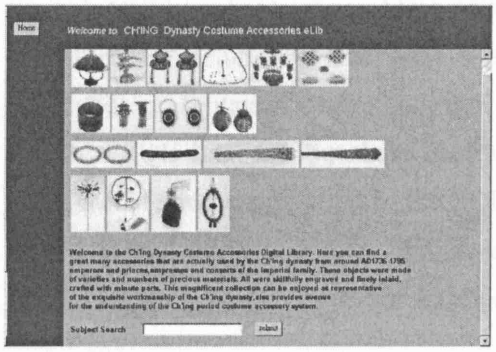

Figure 5: The Ching Digital Image Library

The user interface, as presented in Figure 5, was divided into two parts. The upper part was a thumbnail catalogue (still TIFF images with a resolution of at least 800x1200x24b allowed for further viewing) with eighteen subject groups such as the subject BRACELETS, the subject PENDANTS and so forth. First level viewing presented a sub-catalogue screen view covering objects categorized to the same subject group. The sub-catalogue was dynamic. The user could click on objects of interest to expand them, and this would lead to a second level full screen view with descriptive metadata in it. Below the thumbnail catalogue was a subject search box where the user could enter subject terms. The subject search terms were truncated; that is, the user could enter either COURT or COURT HAIR while searching for objects categorized to the subject COURT HAIR.

The technical requirements for the project were based on Microsoft technology including the Windows 2000 operation system, Microsoft Internet Information Server (IIS) version 5.0 was used as the Web server, Microsoft SQL Server 2000. In addition, the project used the ACDsee 4.0 software (to allow the developers to compress images and give users the capability to zoom in and out for more or less detail) and the XMLSpy 3.5 as an XML documents editor.

Architecture

Microsoft SQL Server 2000 server was used as a back-end database. In creating the database process, three tables (one-to-many relationship) were created representing the Item table, the ItemSubjectlist table and the Subjectlist table. Each object was grouped into one subject group; each subject group was identified alphabetically. To support static XML document output, the

presentation processing was performed either in Web server or client side. XSL, XML and Active Server Page (ASP) residing on the IIS Web server were requested to support static XML outputs in a screen view using DOM and Data Islands (that is, XML embedded in HTML) technologies. To support dynamic XML output in the subject term search box, a model of a typical ASP-based (Database-ADO-ASP-HTML) approach was performed. The difference with the ASP solution at this stage was that the system applied a default XSL stylesheet as client-side presentation design. ActiveX Data Object (ADO) was still used to retrieve the data from the SQL database, but these data were converted to an XML format. This was one of the new XML features supported in the SQL Server 2000. The XML/XSL approach converted the recordset into XML and then applied an XSL transformation to produce HTML. The Ching Digital Image Library's three-tier architecture is shown diagrammatically in Figure 6.

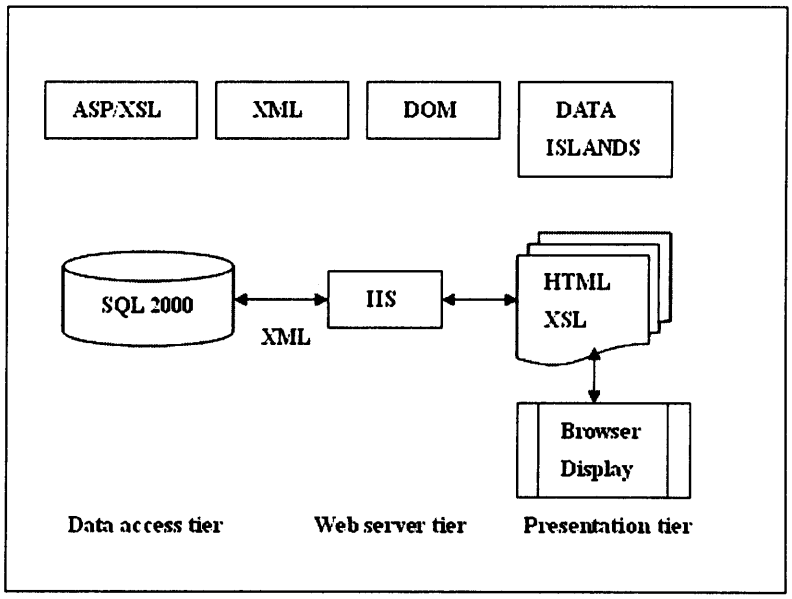

Figure 6: The Ching Digital Image Library's three-tier architecture

Data manipulation

In this section, I use the Ching Digital Image Library as an example to help illustrate data manipulation with Data Islands and XML DOM for static output. In addition, for dynamic output, I will skip the detailed explanation of the ASP page but focus only on the default XSL stylesheet.

Static output:

Method 1: XML Data Islands:

The following section describes the syntax used for embedding Data Island within a page, and

details the object model exposed by the browser to enable the object model to be used. Below is the text of XML source file bl237.xml.

```
<?xml version="1.0" encoding="BIG5" ?>
  <ching>
     <bracelets>
       <title>Rattling bracelets </title>
      <description>outer diameter = 7.05cm inner diameter = 5.75cm
      thickness = 0.85cm  Pair of gold rattling bracelets decorated with flowers in gold
      filigree, kingfisher feather, and inlaid pearls. </description>
     </bracelets>
  </ching>
```

Firstly, the XML file "bl237.xml" will be loaded into an "invisible" Data Island called "bl237" with an ID "bl237" as shown below. I apply HTML IMG element and its attribute SRC to generate the image hosted in IIS Web server.

```
<XML ID="bl237" SRC="bl237.xml"></XML>
<img border="1" src="pic/bl237.JPG" width="636" height="300">
```

Secondly, I bind the Data Island to an HTML page. As the codes below show, I apply DHTML tags <LABEL> to perform the data binding. I use the DATASRC property to specify a binding. The property takes a string that corresponds to the unique identifier of a data source object (DSO) on the page, which is an external XML data source bl237.xml. The string must be prefixed by a number sign (#). When the DATASRC property is applied, the corresponding value of the object will be repeated and displayed in the specified DATAFLD property.

```
<h4><LABEL ID="label_title" DATASRC=#bl237 DATAFLD="title"></LABEL></h4>
<LABEL ID="label_description" DATASRC=#bl237
DATAFLD="description"></LABEL><BR>
```

The resulting HTML page is returned as follows: object image bl237 followed by title and description.

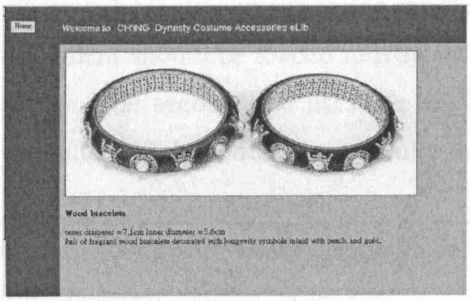

Figure 7: Object image bl237

Method 2: DOM and XSL

XML DOM provides a standard programming interface to a wide variety of applications. With the XML DOM, a programmer can create an XML document, navigate its structure, and add, modify, or delete its elements. In other words, XML DOM provides a pathway to manipulate data stored either in an XML or an HTML file.

I apply the XML DOM technique along with an accompanying XSL stylesheet to demonstrate a method of navigation through XML documents. This is done by accessing on server-side ASP. There are three levels of document structure in the Ching DOM tree structure representing root node, node and childNode. The top root is <ching>, following by a subject element <miscellaneous>, and three childNodes: <picture>, <title>, <description1> and <description 2>. The source file is as follows.

```
<?xml version="1.0" encoding="BIG5" ?>
  <ching>
     <miscellaneous>
        <picture>mc353.jpg</picture>
        <title>Enameled brooch-watch </title>
        <description1>Height = 14.4cm</description1> <description2>Enameled
           brooch-watch in the shape of a butterfly inlaid with jewels.</description2>
     </miscellaneous>
  </ching>
```

Firstly, I start with the method of server-side ASP. Since the HTML page itself does not actually

do anything except display the data, I skip the HTML line. As shown below, the first line of script creates an XML document object with VBScript in an Active Server Page (ASP), followed by the third line of code that tells the default Microsoft XMLDOM parser to load an existing XML document called mc353.xml. Because the XML DOM does not specify how a document should be loaded into an instance of a parser, I use the LOAD method of XMLDOM parser. The second line uses the ASYNC="false" property to tell the parser that it will halt execution until the document is fully loaded.

```
set xml_document=Server.CreateObject("Microsoft.XMLDOM")
xml_document.async="false"
xml_document.load(Server.MapPath("mc353.xml"))
```

Using the same method to create a new XML stylesheet instance called dom.xsl and load the XML file into the XMLDOM parser.

```
set xsl_document=Server.CreateObject("Microsoft.XMLDOM")
xsl_document.async="false"
xsl_document.load(Server.MapPath("dom.xsl"))
```

The last line tells the XMLDOM parser to process the node and its childNodes in the specified XSL stylesheet, and return the result to the browser.

```
Response.Write(xml_document.transformNode(xsl_document))
```

Also, I applied the client-side method, using JavaScript within a HTML page. The two methods worked equally well, and returned the same resulting HTML page which is shown as follows: object image mc353 followed by title and description.

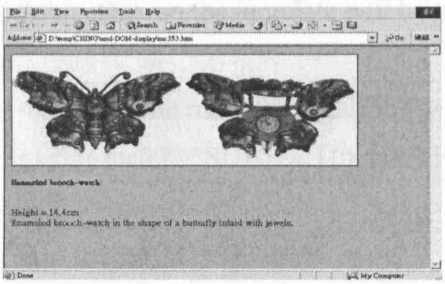

Figure 8: Object image mc353

Dynamic output:

I will not go into the details of ADO but concentrate on the default stylesheet for the dynamic output. For dynamic output, I experimented with different version of XML source files, which are in upper case. I defined a template for a whole branch of the XML document. The <xsl:apply-templates> element and SELECT attribute tell the template to process the ITEM element wherever it appears within the XML document, and specify in which order the child nodes are to be processed and sorted as output elements. The template will output starting with TITLE child node, followed by ID, SUBJECT, DESCRIPTION and ultimately the PICTURE child node.

```
<xsl:template match="/">
<xsl:apply-templates select="//item" order-by="+title"/>
```

The <xsl:value-of> element is used here to select the value of the XML element and add it to the output stream of the transformation. Note that I apply new templates for SUBJECT and PICTURE child nodes, so their value will be abstracted under new required attributes. I will explain this in the following two block codes.

```
<TD WIDTH="140"> <xsl:value-of select="TITLE" />
<TD WIDTH="60"> <xsl:value-of select="ID" />
<TD WIDTH="120"> <xsl:apply-templates select="SUBJECT" />
<TD WIDTH="180"> <xsl:value-of select="DESCRIPTION" />
<TD WIDTH="160"> <xsl:apply-templates select="PICTURE" />
```

For the SUBJECT child node, the conditional <xsl:if> element will be applied unless a specified condition is matched. I use a predicative expression (written within square brackets) to test the condition and select the subset of the nodes based on the test. The value of the required MATCH attribute will be evaluated. In Ching, each subject in the Subjectlist table is assigned a SubjectNo, so the template will be applied only if the subject number is not equal to zero.

```
<xsl:template match="SUBJECT">
<xsl:if match="SUBJECT[index() != 0]">
```

For the PICTURE child node, I use the <xsl:attribute> element to add attributes to IMG elements. Again, I apply an IMG tag and SRC attribute to retrieve and display images as the value of PICTURE.

```
<IMG WIDTH="160" HEIGHT="150">
<xsl:template match="PICTURE">
<xsl:attribute name="SRC">
 jpg/ <xsl:value-of />
```

The result of the transformation will look like this: all the items categorized to the search subject group are retrieved and displayed as the table below.

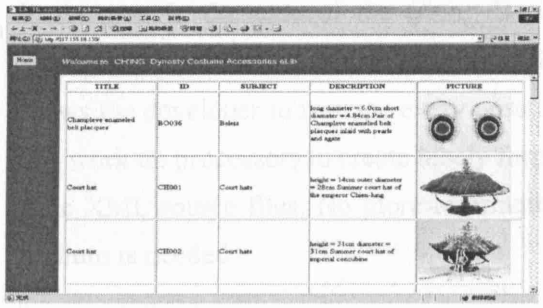

Figure 9: A dynamic search example

Advantages of XML solution

XML is simply a textual markup language for describing the structure of application data. Using XML to send data to the browser opens up many more possibilities for Web applications [Howlett and Dunmall, 2000].

Firstly, the use of XML is a key to the three-tier architecture [Sall, 1998]. By breaking down an application into three distinct and separate tiers, the presentation tier, the Web server tier and the data access tier, much of the presentation processing on the server side has been shifted to the client side and can reduce network and Internet traffic, making the Web faster [Lander, 1997; Freter, 1998b; Ragnarsdottir, 1999]. The developer can manipulate the source XML data on the client, and that eases the work of the server. In the Ching Digital Image Library, at the data tier, relational data are transformed into XML documents. At the Web server tier, XML/XSL documents are requested and rules are executed; at the presentation tier, XML documents are transformed by an accompanying XSL stylesheet based on the user's needs and are sent back to the screen.

Secondly, one fundamental characteristic of XML is the separation of the structure, content and presentation of data. In a digital library, this separation of data from presentation in XML makes it possible for the integration of data from diverse sources.

Thirdly, a developer can use the XML format for extensible and flexible data storage, presentation and interchange. For example, we can extend the source XML document <description> element to <description1> and <description2> in the DOM method with new tags representing new data structures and processing rules. As XSL is built with presentation flexibility, we can create a number of stylesheets for different scenarios.

Fourthly, the XML approach allows the developers to reuse the same XML source document. On the other hand, the same XSL stylesheet could be shared by different XML source documents. In the case of the Ching Digital Image Library, one default XSL stylesheet is responsible for all the database objects while performing a dynamic search. Additionally, XSL allows the developer to manage every possible formatting option. For example, a developer can also work on processors to create nicely formatted PDF documents for a print medium using the same XML source files. No more reformatting of the same content to be presented in another medium is needed.

Ultimately, the Data Island and DOM methods described above share the same advantages – efficiency. The combination of four Web technologies, HTML, DHTML, XML, XSL, allows for rich content to be generated with only 5 to 8 lines of code. Data binding approaches ensure that the providing of the data and the consuming of HTML elements in the display are simultaneous. The potential disadvantage to using this method is that it would be problematic if there were more than one bound element to a tabular data consumer. The DOM represents a tree view of the XML document, thus mirroring closely XML document structures. Data binding manipulates data that have been serialized as XML in a way that is more natural than using the DOM. For example, the developer finds it easier to use Title, Subject, and Description in the data binding approach rather than Element, Attribute in the DOM approach.

4.8 Relationship of XML to Databases

XML itself is not a database but the XML family provides many though not all of the components found in databases. XML provides storage of documents, DTDs and schema languages, query languages, for example XQuery and XML-QL, and programming interfaces such as DOM and SAX (Simple API for XML). XML coupled with a database gives greater power than the sum of the parts in a Web application. From these practical examples based on the Ching Digital Image Library, we can understand that the potential impact of XML is promising. XML provides a standard syntax for exchanging data; thus, Web servers and

applications encoding their data in XML can quickly make their information available in a simple and usable format, allowing a degree of data and application integration.

There are instances of the implementation of relational databases in digital library projects [Columbia University Digital Library Project, 1998; Staples and Wayland, 2000; Rhyno, 2002a; Johnston, 2005]. Digital libraries of the future will provide access to a wealth of information media. Interoperability of data models, of databases, of system, and of applications will be crucial. Particularly, the representation of XML on standard multimedia is still at its early stage. As we have seen, research into applications on XML-related technologies and XML-oriented tools are substantially active and ongoing. Database-style technology applied to XML could play an important role in both of them. As discussed in Section 6 of this Chapter, the native XML database environment is immature and is not used as much as XML-enabled databases. One main reason for this could be the market needs; because of the success of SQL, most institutions have invested their data in relational databases. The other possible reason could be that the entire XML application environment is still unclear and this is difficult for the development of XML query languages.

On the other hand, we see that a large proportion of data in many libraries is stored in relational databases. This is cost-effective in terms of leveraging an organization's existing investment and expertise in relational database management systems. The leading enterprise relational database vendors will focus on providing integrated XML support to their database platforms. In this sense, relational databases with object-oriented extension database management systems are likely to be the potential candidates to store and manage XML documents in digital libraries. In the instance of my case studies, they all use relational databases in their system infrastructures, Perseus uses PostgreSQL and MySQL; Michigan and the Library of Congress use Oracle, along with a search engine which supports object behaviour. This will be shown in my three case studies of digital library initiatives discussed in Section 2.3 of Chapter 8.

Chapter 5

XML and Metadata Standards and Interoperability

In this Chapter, I discuss the definition of metadata and the importance of metadata in the administrative functions that enable digital libraries to operate. In the context of metadata systems for the Web, XML is likely to be a promising technology. Metadata is a key component for a broad range of applications that are emerging on the Web. In this Chapter, I also review a range of leading metadata efforts in XML formats which are used in digital library projects. This includes both the metadata standards and the standard infrastructures being developed to support them.

5.1 Metadata

According to Shelley and Johnson [1995], metadata means "data about data"; Baker and Lynch [1998] defined it as an "Internet-age term for structured data about data". The term metadata is used differently in different environments [Hodge, 2001]. For example, in Section 3.1 of this Chapter, when discussing metadata in the Resource Description Framework, this primarily refers to "machine-understandable information about Web resources or other things" [Berners-Lee, 1997]. In the library community, metadata is often used to indicate description information contained in abstracting and indexing information and catalogue records as a means of retrieval information. In the past, libraries used catalogue cards and printed bibliographies and these were their metadata systems, though that term was never used then. Since the 1960s, MARC cataloguing has been playing an important role as metadata in the mechanism for sharing catalogue records in the online public access catalogue (OPAC), yet more complex requirements are demanded for metadata to facilitate the management and use of networked resources [Baca, 1998].

The primary reason for digitizing collections is to increase access to the resources held by institutions. Implementing associated metadata is a way to maximize access to broad user communities. Such metadata will not only provide granularity of description of collections, but also support long-term management of objects in collections.

Lagoze and Payette [2000] proposed four functional uses of metadata for networked information resources on the Internet:

- **For resource discovery** Digital resources that have been stored in computer systems or on computer readable media for current and future use will need to be retrieved as reliably as possible for as long as possible. In this sense, the storage systems will have to incorporate systems with strength in resource discovery such as RDF.
- **For presentation and navigation** Digital resources not only should be identified and located, but also be browsed and navigated. Structural metadata represents the relationships between digital objects and their component parts. It may support many functions including key access points, browsing, navigation and structural relationships.
- **For rights management and access control** Metadata issues in electronic rights management have been carefully considered as a consequence of the growing importance of copyright protection in the digital environment. Rights management in a digital preservation context is crucial. Iannella [2001] discussed how, in a digital rights management functional architecture, the descriptive metadata and rights metadata covering elements such as Parties, Rights, Payments, are stored in a repository which enables the access/retrieval of content in potentially distributed databases and the access/retrieval of metadata. XML/RDF based DOI (Digital Object Identifier), which is discussed below, supports a system for identifying and exchanging intellectual property in the digital environment and is an example of such.
- **For administration and preservation** Information associated with the functions of managing and administering information resources, and ensuring their long-term preservation.

Also, I add to Lagoze and Payette's four functions; the function of confirmation of identity once a resource has been discovered. For example, metadata can be used for uniquely identifying an item. The standard numbering systems serve this function such as DOI, ISBN (International Standard Book Number) and ISSN (International Standard Serial Number). ISBN and ISSN were developed for recording printed material in automated systems. They have since been applied to digitized books and serials. On the other hand, DOI has been developed for the digital world [DOI, 2002]. The digital library community regard the DOI system as one of the most important digital library technologies, as it provides a solution for identifying and exchanging intellectual property in a digital environment [Haigh, 1998]. The primary purposes of DOI are effective rights management and digital commerce. Many publishers have begun implementing the DOI system such as Wiley and Academic Press according to a Wiley press release [John Wiley & Sons Ltd., 1999].

An early successful example of the full-scale DOI implementation is CrossRef [CrossRef, 2003], a scholarly publisher collaborative launched in 1999, where scientific and other scholarly professional articles are assigned a DOI on publication that enables linkage from citation to source. As of 2005, more than three million article records are available to CrossRef members and affiliates from the library and information communities. For example, libraries using the Ex Libris library management system can utilize DOI and CrossRef linking in a fully integrated way. In addition, one of the significant achievements of DOI is the practical implementation of DOIs with related initiatives such as the OpenURL framework in contextual linking [CrossRef, 2003]. OpenURL is the syntax for transporting metadata and identifiers within URLs. CrossRef uses the OpenURL syntax in its own system, allowing library participants to retrieve publisher metadata from the CrossRef system for the purposes of article-level linking to local holdings. The important advantage of these linking mechanisms is that library users or researchers in library environments may navigate online journals via DOI-based citation links through CrossRef, and then redirect to the publisher's full-text source.

One could also say that metadata itself is data; that is, metadata can be stored in exactly the same way as data and it can be stored like data in a resource. Therefore, the distinction between "data" and "metadata" is not an absolute one; the distinction is created primarily by a particular application ("one application's metadata is another application's data") [Lassila, 1997]. Hence, one could even say, metadata can describe metadata; that is, metadata itself may have attributes. For example, price lists which are metadata have their own administrative metadata, such as expiry dates, and so this is metadata about metadata. Therefore, metadata is data and it can have other data about itself. These recursive concepts predate electronic systems: there have always been bibliographies of bibliographies and catalogues of catalogues.

Even specialists in the field when thinking about metadata are often considering the metadata that is embedded in the markup on a Web page. However, not all metadata is available in this way. During information search and retrieval activities in databases on Web servers, Web pages are often generated dynamically. Although the popularity and utility of the Internet and the Web have begun to reveal the power of digital libraries, traditional search engines cannot retrieve precisely the information needed by users nor the content in the deep Web, where Web pages do not exist until they are created dynamically as the result of a specific search on a database which is not indexed by Web search engines [Bergman, 2001]. Examples include information stored in tables created by relational databases which is accessible only by query, or information that is dynamically changing in content like multimedia files. Bergman's study observed that the deep Web is of a higher quality than the surface Web and that topic databases made up more than half of the deep Websites, and the deep Web is growing faster than the surface Web. One of the most

important findings from Bergman is that there is a lack of awareness that these kinds of data with their critical contents even exist. However, this kind of system can be used to produce any kind of output which can include data in any of the formats which will be discussed later such as TEI.

There are other kind of metadata which can be used in digital libraries. Classification systems are types of controlled indexing vocabulary which have long been implemented as part of standard cataloguing practice in libraries and museums and are therefore important in digital libraries. There is a broad range of systems in existence both for subject domains such as Medical Subject Headings (MeSH), or for general purposes, such as Library of Congress Classification (LCC) and the Universal Decimal Classification (UDC). These controlled vocabularies are now being applied to resource discovery on the Web via thematic keywords in metadata resource descriptors, which could possibly provide support in information search and retrieval [Tudhope and Cunliffe, 1999]. Through semantic index technology, links between concepts in the subject domain can be expressed by the semantic relationships in these vocabulary systems [Koch, 1999; Broughton, 2001; Edinburgh Engineering Virtual Library, 2005].

5.2 Metadata Standards

To fulfil the functions of metadata described above and to do this successfully, metadata created for digital libraries should meet certain requirements. These include flexibility, interoperability, easy input, easy searching and browsing, being able to deal with diverse digital objects, being accessible outside the institution and allowing for multiple points of access to the collections [Lee, 2001, pp.103-109].

To achieve these functions, it is important to develop metadata strategies, including well-conceived data models, well-structured metadata, standards for expressing and exchanging metadata and common semantics for metadata. Therefore, an underlying system will be needed, that is, both flexible and extensible enough to cope with the complex metadata and meet the metadata functions over time.

There are several metadata schemas available provided by various subject metadata communities that have the potential of being widely used as standards for the description of materials in digital libraries: for universal descriptors, there are the MARC formats and Dublin

Core Metadata; for text descriptors, there are TEI Headers; for images and objects, there is the Consortium for the Computer Interchange of Museum Information (CIMI); for mixed collection and item level descriptions, there is the Encoded Archival Description (EAD).

5.2.1 Dublin Core

Dublin Core (ISO 15836:2003(E)) is a set of fifteen metadata elements designed specifically to describe and identify the majority of types of resource available on the Internet. The Dublin Core Metadata Initiative (DCMI) began at a meeting held at OCLC headquarters in Dublin, Ohio in 1995, hence the name, with participation from the National Center for Supercomputing Applications (NCSA) [OCLC, 1995]. The simple and general Dublin Core description features have proved to be a useful scheme when intended to retrieve information from heterogeneous descriptions [Wert and Hernandez, 2001]. The name "Core" indicates an assumption that Dublin Core will coexist with other metadata sets and will be a kind of core or lowest common denominator to other formats. Dublin Core is engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. Additionally, in order to meet the increasing demands of the complex relationships which exist within many digital resources, the Dublin Core Metadata Initiative has extended its use by developing a batch of fifty-two qualifiers. These are referred to as qualified Dublin Core [DCQ, 2000]. Dublin Core has been working actively alongside other initiatives and markup systems [Weibel and Koch, 2000], for example IEEE Learning Technology Standards Committee (LTSC), Global Information Locator Service (GILS), MPEG-7 and many others.

XML is notable for its flexibility. Tags can be added to any extent desired. Dublin Core is a metadata system known for its flexibility though not always commended for this flexibility, as we shall observe. It is not surprising that Dublin Core has been used with XML or that qualified Dublin Core has been used as well, giving different kinds of sophistication and depths of definition. In this section, I look at a number of projects that have used Dublin Core for different purposes. However, the review of existing Dublin Core community practice is still ongoing. One argument is that qualified Dublin Core may provide the granularity but not necessarily completely. To give one example, comparing qualified Dublin Core with MARC: each distinct Library of Congress Subject Heading in MARC is categorized as being a topical, personal, corporate or meeting subject heading, but qualified Dublin Core does not make such deep distinctions.

Baker [2000] views Dublin Core as a language for making a particular class of statements about resources, especially when it works with RDF/XML. OCLC has been actively involved in the

implementations of RDF/XML. Among the fruitful efforts, the DCMI Open Metadata Registry (OMR) is a prototype for the DCMI registry [Weibel and Koch, 2000]. A metadata registry is a term that refers to a formal system that records the semantics, structure and interchange formats of any type of data, for example, data found in databases, messages, documents and other applications [Lawrence Berkeley National Laboratory, 1997]. OMR is expected to become the definitive repository of DCMI schemas, making it a central management tool for the Dublin Core and associated elements or qualifiers. OMR also plays a role as a central component for managing relationships among various metadata communities which could make efficient the agreement on standard representations of schemas from different communities [Weibel and Koch, 2000].

A similar project is the Development of a European Service for Information on Research and Education (DESIRE), which was an effort of UK Office of Library and information Network (UKOLN) in the area of metadata registry [DESIRE, 1999]. Heery and Patel [2000] at UKOLN continued the experiment in metadata schemas and registries in the DESIRE project, and developed the RDF/XML-based Application Profiles prototypes to explore the way the Dublin Core Metadata Element Set and other metadata standards were used in the real world. The work on Application Profiles has led to the combining of Dublin Core with other metadata element sets, and thus supports richer descriptions drawn from different metadata communities. The JISC Information Environment Metadata Schema Registry (IEMSR) can be regarded as a continuation of UKOLN's effort in the area of metadata registries, complementing earlier research outcomes to provide a metadata schema registry service for the JISC Information Environment [Johnston, 2005].

Another application of Dublin Core and RDF/XML is the DCMI Collection Description working group. The group has been addressing the Collection-Level Description (CLD) issue in collaboration with UKOLN through the UK Research Support Library Programme (RSLP) [Weibel and Koch, 2000]. The RSLP Collection Description project enables project partners to describe their collections in a consistent and machine-readable way. The description of collections at such a shared level of granularity has been recognized as significantly contributing to the cross-domain work shared by libraries, archives and museums. Continuing attention has been paid to CLD since it was first developed [Johnston, 2002].

The MANTIS project is another practical effort from OCLC. The project employs XML/RDF techniques for handling metadata resource descriptions, distributed searching and transforming results for flexible display. This effort aims at lowering the barriers to the acceptance, distribution and use of the tools by developing a toolkit for building Web-based cataloguing

systems [Shafer, 1998]. MANTIS has been used as a basis for several OCLC projects such as the Cooperative Online Resource Catalog (CORC) project, and the next generation OCLC SiteSearch Image Support Package.

Digital libraries should possess the ability to store and deliver complex multimedia resources that combine text, image, audio and video components. The Harmony project developed a "framework to deal with the challenge of describing networked collections of highly complex and mixed-media digital objects". The work studied mechanisms for expressing the conceptual model bringing together work on the RDF, XML, Dublin Core and MPEG-7, and retained "the focus on the problem of allowing multiple communities to define overlapping descriptive vocabularies for annotating multimedia content" [Harmony, n.d.].

It is important that data produced from the reporting of research initiatives in papers in traditional and electronic form are preserved in a way that will enable them to last for posterity. One of many examples of such material is D-Lib Magazine, which publishes articles about digital library innovation and research. It has adopted as its metadata method DOI with an associated file which contains simple metadata for each article, and uses XML/Dublin Core as its metadata system [Arms, W.Y., 1999].

The Contemporary culture Virtual Archive in XML (COVAX) project addressed the issue of access to cross-domain (archives, libraries and museums) information via XML DTDs [COVAX, 2001]. It reflects the increasing awareness of XML as a Web standard for building a digital library of the future, which is multilingual, multicultural and sustainable. COVAX performed as a distributed database which was accessed as a single one, and acted as a meta-search engine, offering access to book references, finding aids, facsimile images, museum items and so on. It was decided to use elements of the Dublin Core Metadata Element Set as access points in a common format, and conversions were made between that and the different DTDs used in COVAX. This common format made it possible to make information available through search and retrieval from the pertinent elements of each DTD in use.

One implementation of Dublin Core in the digital library environment is Open Archive Initiative (OAI) [OAI, n.d.]. OAI is an international initiative that attempts to combine the best of library and Internet techniques into a new model for accessing scholarly resources. While the OAI has its roots in the scientific e-print community, the library community, with support from the Andrew W. Mellon Foundation, has generalized the concept into a universal model for research metadata harvesting [OAI, n.d.].

The current OAI technical infrastructure, which is specified in the Open Archives Metadata Harvesting Protocol, defines a mechanism for data providers to expose their metadata through an HTTP-based protocol. OAI has adopted Dublin Core as a preferred metadata schema though others may be used alongside. One of the disadvantages of Dublin Core in OAI is that if only the minimum of unqualified Dublin Core is used, there is no granularity. On the other hand, OAI has recommended XML as a common protocol to provide access to data across hardware platforms and operating systems and to describe information in granular detail if the metadata schema supports it, benefiting precision searching. The OAI model has made possible a wider range of digital resources held in many individual systems worldwide to central collections that are of academic and scholarly interest. In Section 2.1.1.3 of Chapter 8, I will discuss how my three case studies deploy OAI in their digital library infrastructures, and the disadvantages of OAI.

The Dublin Core initiative has laid the foundation for global, interdisciplinary resource discovery that promises to improve the global networking. However, deploying qualified Dublin Core will continue to be an experimental endeavour for more time to come before it is widely accepted by digital library initiatives. Indeed, many projects will continue to use Dublin Core alongside other metadata systems, or to use Dublin Core as a common format. Solutions on how to cope with issues such as extensions, interoperability and the like will emerge from further work on underlying models and from the practical experience gained from projects with real-world problems to solve.

5.2.2 TEI Header

In Section 2.1.1 of Chapter 3, I discussed the benefits of electronic text in many aspects. The TEI Header is one serious attempt, and probably the first attempt, to define a structure for metadata for electronic texts [Hockey, 2000, Chapter 3]. This is needed because it will be useful for research purposes that metadata for electronic texts contains information about the encoding principles and source from which it was digitized such as author and title, and information about any revision made to the original texts. Renear [1997] pointed out that the TEI researchers and TEI text-based projects had brought a wide variety of disciplinary methods and problems, contributing to the development of encoding techniques and theories. For example, historical documents would be enhanced by links to biographies; literary works could be enhanced by links to etymological dictionaries. The wide variety of problems reflected the wide range of applications ranging from manuscripts to modern texts, dictionaries to manuals. Robinson et al. [1999] also thought that the acceptance of TEI in defining encodings for Humanities texts and the recognition that the combination of online manuscript descriptions and digital images of

manuscripts could vastly increase access by scholars and others to the manuscripts, and their study fueled interest in the development of an agreed standard for the encoding of machine-readable manuscript descriptions.

The TEI Header is the metadata part of the TEI Guidelines, which have been endorsed and are used by leading text centres in the United States and Europe. Digital Library Federation promotes the adoption of TEI as a standard and best practice for digital libraries. Since TEI is adopted in my three case study digital libraries, the following section is a small discussion of TEI Header structure.

As far as the metadata is concerned, the TEI Guidelines specify that every TEI text must be preceded by a TEI Header that describes the text. The TEI Header can be used in different operational contexts. It can exist as part of a conformant text; independent headers can be used in catalogues or databases to refer to a remote TEI encoded text; also, it could be used as metadata to describe networked resources, mostly text information but, arguably, any type of networked resources. I will discuss further on the problems of the TEI Guidelines in Section 2.1.1.2 of Chapter 8.

The TEI Header, tagged <teiHeader>, is a mandatory element in a TEI document. It has four major parts which form the syntax of a TEI document reproduced below, of which the file description is the most important [Sperberg-McQueen and Burnard, 1999, Chapter 5]:

- a file description, tagged <fileDesc>, containing a full bibliographical description of an electronic file. The <fileDesc> element contains three mandatory elements:

<titleStmt>	groups information about the title of a work and those responsible for its intellectual content.
<publicationStmt>	groups information concerning the publication or distribution of an electronic or other text.
<sourceDesc>	supplies a bibliographic description of the copy text(s) from which an electronic text was derived or generated.

The <fileDesc> element is the sole element required in a TEI Header, while the others are optional; thus, the following structure is a minimal file description:

```
<teiHeader>
  <fileDesc>
   <titleStmt> ... </titleStmt>
   <publicationStmt> ... </publicationStmt>
   <sourceDesc> ... </sourceDesc>
   </fileDesc>
</teiHeader>
```

• an encoding description, tagged <encodingDesc>, documents the relationship between an electronic text and the source or sources from which it was derived. It covers nine optional subdivisions. One of these, <projectDesc>, describes in detail the aim or purpose for which an electronic file was encoded, together with any other relevant information concerning the process by which it was assembled or collected.

• a text profile, tagged <profileDesc>, contains classificatory and contextual information about the text, such as its subject matter. The core <profileDesc> element has three optional components, represented by the following elements: <creation>, <langUsage>, <textClass>. The <textClass> groups information which describes the nature or topic of a text in terms of a standard classification scheme, such as the Universal Decimal Classification, the Library of Congress Classification, or any other system widely used in library and documentation work. These are particularly important as a means of documenting the organizing principles of corpora or collections.

• a revision history, tagged <revisionDesc>, which allows the encoder to provide a history of changes made during the development of the electronic text. It is important for version control and for resolving questions about the history of a file. Revision history contains tags like <date>, <respStmt> and <item>.

A TEI Header can be a complex object, or it may be a simple one. The amount of encoding in a header will depend on the applications. Some application areas will require more specialized and detailed information than others. The Guidelines therefore define not only a core set of elements, but also additional tag sets, which may be called into use as extensions, as and when needed [Sperberg-McQueen and Burnard, 1999, Chapter 5]. The Oxford Text Archive found the TEI Header an invaluable tool and used it as a means of managing its large collection of electronic texts [Morrison et al., 2000]. Furthermore, TEI Header metadata with its rich tag sets can be the source of new metadata sets, or be mapped to library cataloguing MARC records, Dublin Core element set and RDF.

TEI Header metadata has been widely adopted in many leading projects and institutions

worldwide. One example of the TEI Header used with XML is the Walt Whitman Hypertext Archive (Whitman's major work: *Leaves of Grass*). The project was supported by XML technology, presenting easy and convenient access to scholars, students and general readers as an electronic research and teaching tool [Folsom and Price, 2000].

TEI Guidelines provide users with a fixed set of tags and attributes which can be easily extended, but do not provide users with any rules for determining the values of the tags and attributes. In other words, the Guidelines provide no "authority control" for the form of names, subjects, uniform titles or serial titles [Giordano, 1994]. Applications of each discipline and even different information professions (such as archivists and librarians) working in those disciplines can apply their own "languages" in the values they give to the tags and attributes. This limits the TEI Header's ability to provide fine-grained retrieval, since retrieval often takes place over cross-disciplinary data, and tags and attributes are unlikely to have been given consistent value across these different disciplines. Furthermore, TEI Guidelines by their very nature make sense only in the context of text objects. This could be a disadvantage when building and maintaining a heterogeneous digital library. I will discuss further on the problems of TEI Header in Section 2.1.1.2 of Chapter 8.

In general, since SGML/TEI has been implemented by the major digital library encoded-text programmes worldwide, XML/TEI combination will continue to further facilitate development in the metadata used in managing digital library services and collections.

The TEI Header is the main technology in my case studies, and the Dublin Core is the metadata which is currently most under discussion. In the following paragraphs, I compare the two mechanisms in terms of their implementation in digital libraries.

The TEI Header may be large and complex while Dublin Core is usually small and simple. They are both extensible. Simple Dublin Core records can be used as a starting point for the creation of more complex descriptions while a TEI Header can be created and coexist with a new independent header. The simplicity of Dublin Core lowers the cost of creating metadata and promotes interoperability at a technical level, if not at the intellectual level. The complexity of the TEI Header suits well the needs of Humanities and linguistics communities. Conversely, the simplicity of Dublin Core does not accommodate the semantic and functional richness supported by complex metadata schemes while the complexity of the TEI Header is not intended for a non-Humanities digital library. Dublin Core has a more active and versatile working group, in particular, it has a Libraries Working Group while the TEI Header has been limited to the Humanities and linguistics communities.

The TEI Header has its roots in Humanities computing projects worldwide but is struggling to keep up with the fast changing networked world which I will discuss further in Section 2.1.1.2 of Chapter 8. Dublin Core is recognized by general purpose digital libraries but is struggling to achieve a level of standardization that might have been expected, since there are many different ways of formulating the data elements. It is, however, not suitable to judge which metadata schema is the best in practice for a digital library, but they would be adopted by an institution according to the functions and needs of its digital library.

Although Dublin Core does not have such large-scale deployment in digital libraries as has the TEI Header, the use of Dublin Core is increasing. There remain complex dimensions to the development of qualified Dublin Core such as how to support the richer semantics that qualified Dublin Core is intended for. As we have seen in this Chapter, active working groups and deployment projects are experimenting with more sophisticated deployments. As time goes on, I believe that these efforts are likely to mature; the digital library community will most probably decide for itself whether this is the way to go and whether to develop qualified Dublin Core sets for its own purposes.

5.2.3 Encoded Archival Description (EAD)

The Encoded Archival Description (EAD) initiative is an XML DTD (originally SGML) developed as a standard for encoding finding aids for searching and displaying on the Internet. It has been developed by the Society of American Archivists (SAA), but it is maintained by the NDMSO of the Library of Congress, which acts as a maintenance agency for many standards in the bibliographic and archival field. The definitions of the tags are based on the data description elements of the General International Standard for Archival Description (ISAD(G)) developed by the International Council of Archives. That is a standard equivalent to IFLA's ISBD International Standard Bibliographic Description used widely in the library world and for instance as the underlying rules on which Anglo-American Cataloguing Rules (AACR) are developed. In 1998, the EAD Working Group revised the beta tag library to be compatible with XML in order to embrace the future application and the evolution of Web browsers and protocols [Library of Congress, 2003a]. EAD is a non-proprietary open data structure; it defines and controls the structure of archive finding aids, and facilitates archive finding aids to be delivered on the Internet.

The EAD design principle is heavily influenced by TEI [Library of Congress, 1998c]. For example, the EAD data model includes a finding aid header which is similar to the TEI Header, and TEI naming conventions and tag structures are utilized. EAD DTD contains three high-level

elements:

- <eadheader> used to document archival description or finding aid.
- <frontmatter> used to supply publishing information such as a title page.
- <archdesc> contains the archival description itself.

The document structure of archive finding aids is hierarchical where nested levels of description are used to describe the content of the archive in a whole to part relationship. This seems to map well to XML. EAD is used only in the archive community. There were attempts to develop archive databases using MARC and the Library of Congress developed USMARC for Archival and Manuscript Collections (MARC AMC) [Smiraglia, 1990]. An attempt was made in the United Kingdom to develop something similar based on UK MARC [Ray, 1994]. Archivists have always found it difficult to produce a flat file in the form of a MARC record which at the same time does justice to the relationship of parts to whole. Whereas MARC is a codification of the catalogue card, EAD is a codification of the archivist's archival list, the equivalent of a librarian's catalogue in terms of its use as a finding aid.

Many EAD XML applications have been seen in the academic, research, government or private sectors, for example the EAD/XML Finding Aids project at Cornell Institute for Digital Collections, which since 1999 has been experimenting with delivering archival finding aids encoded in EAD in XML, making use of XSL [Cornell Institute for Digital Collections, n.d.].

Traditionally, a written work began as a handwritten manuscript, and in many cases this handwritten material is valuable to researchers. The number of modern manuscripts and letters stored in European institutions is enormous; yet, the access conditions are not equally well developed compared to traditional books. The aim of the EU funded project MALVINE (Manuscripts And Letters Via Integrated Networks in Europe) is to facilitate access to modern manuscript holdings kept and catalogued in European libraries, archives, documentation centres and museums [MALVINE, 2003]. A converting technique was developed that translated catalogues held in various local formats into a standard interchange format in XML using EAD. The converting tool demonstrates the translation from the EAD format into various native formats using XSL stylesheets techniques.

5.2.4 Computer Interchange of Museum Information (CIMI)

In order to perform the basic management responsibility for preservation of museum information and integration of museum functions well, the Museum Computer Network (MCN)

launched its initiative for Computer Interchange of Museum Information (CIMI) to develop standards that could support museum data management requirements. The CIMI framework adopted the TEI framework as its design principle, providing guidelines for museums, museum consortia and vendors of museum services, and defining the purposes and contents of specific exchanges of data.

The CIMI project Cultural Heritage Information Online (CHIO) took TEI Lite (a DTD that includes only a subset of the whole TEI system) as its starting point to support CHIO requirements for online access. CHIO found the advantage of TEI Lite was that the smallest framework was well able to support the tagging of standard features of exhibition catalogues, and to provide support for a number of concepts that are of direct museum significance [Light, 1996].

CIMI with experience in developing, testing and implementing standards, and its expertise in developing community skills and awareness, had been dedicating itself to encourage the use of standards through further research and projects, disseminating the use of standards to the museum community in order to take advantage of network services. CIMI found that the system-neutral approach of XML was a flexible means of describing museum content which allowed different systems to manage and provide access to museum knowledge. Having an XML-DTD would help museums to tackle the practical difficulties of implementing content standards across increasingly complex systems architectures in place within organizations. The CIMI organization planned an XML-DTD testbed project with testing by the CIMI membership of the first full version of the SPECTRUM XML-DTD [CIMI, 2003]. Unfortunately, CIMI got into financial difficulty during 2003 and disbanded at the end of that year, and now the project has been taken over by the British MDA (Museum Documentation Association) who are the owner of the SPECTRUM standard [CIMI, 2003].

The ability of the SPECTRUM XML-DTD to interoperate with the CIMI, TEI and EAD DTDs will be tested, and the potential for using the TEI DTD as a means of updating the CIMI SGML DTD will be investigated [CIMI, 2003]. This will result in an increase of the museum community's understanding of how they can make best use of XML [CIMI, 2003]. CIMI's XML programme highlights an increased awareness of the potential of XML within the museum community and the wider cultural heritage sector.

5.2.5 MAchine-Readable Cataloging (MARC)

The purpose of MARC is the exchange of bibliographic data, and most MARC-based systems

import MARC records into their own proprietary structure and convert data there into the different representations required for index building and the display of information in its many and varied formats [Gredley and Hopkinson, 1990, pp.24-31]. MARC programs will not be able to interpret other structures and MARC will not be understood by any other than these programs designed for the purpose.

MARC bibliographic records exist for a large percentage of the materials available or potentially available through digital libraries at least for the materials in the original printed form where the digital item is the same intellectual object. Therefore, existing bibliographic information may be captured, so that these materials do not have to be recatalogued, though amendments will be made to the records to indicate that, in this instance, they refer to a digitized version. Two of my case studies did exactly this. The main use of MARC metadata in the digital library will be for importing MARC bibliographic records from existing MARC-based systems. Additionally, many electronic resources come with their own metadata which can also be captured, but in order to make these available to users in an integrated manner in a catalogue of a library which is based on MARC records (as most are), they will have to be converted to the MARC format.

There are also efforts in mapping MARC metadata to other leading metadata such as Dublin Core. MARC is a more specific format, and so can provide a Dublin Core record quite easily, even a qualified Dublin Core record [NDMSO, 2001]. The problem lies in converting from Dublin Core to MARC, which libraries need to do if they want to harvest the Dublin Core for their traditional library systems. Even after the conversion of a qualified Dublin Core record, it is likely that there would be manual intervention required to bring that harvested record up to an acceptable standard for a traditional library catalogue. Therefore, many libraries are setting up separate databases for cataloguing their digital library materials rather than putting that extra effort into the cataloguing. A similar problem appears in the mapping between MARC and TEI Header, which I will discuss along with my case studies in Section 2.1.1.2 of Chapter 8.

Since SGML or XML can accommodate complex bibliographic data, it appears possible to incorporate the complex specificity of the data definition of MARC into the more universal format of XML. The earliest effort started from 1995-1998 at the Library of Congress with a literal mapping of MARC to SGML (and later XML in 2001). The Network Development and MARC Standards Office (NDMSO) worked initially in 1995 on an SGML DTD for MARC. It is an extensive DTD because it is element-based. NDMSO provides the library and information community MARC SGML DTDs conversion utilities to support converting data between TEI Header and MARC records [NDMSO, 2003].

Medlane is a more recent effort in addressing MARC to XML research and it is also the basis of one of the books on the subject of XML in libraries discussed in Section 2.2 of Chapter 1 [Miller and Clarke, 2003]. The Medlane project at the Lane Medical Library at Stanford University has involved converting catalogue records to XML for integration with other Web resources. They released Java client-server conversion XML MARC software with source code in December 1999, and developed sample DTDs to explore restructuring and simplifying MARC. After the release of XMLMARC software, Medlane has moved its research to investigate the match of the complex data structure of MARC bibliographic records to the most appropriate data storage model. Medlane has been encouraged by the ease of working with native XML databases, in particular, dbXML. However, as I discussed in Section 6 of Chapter 4, the XML-enabled databases have a more mature data model and tools available in the market than native XML databases. Medlane opted to store their MARC data in Oracle CLOBs, but to await the development of a mature and robust native XML database infrastructure [MEDLANE, 2002]. The Medlane case may be the method that most digital libraries have used to store their data.

The library community in general and the MARC community led by the Library of Congress in particular are conscious of the need to ensure standard practices. At the 2003 meeting of the International Organization for Standardization, the Danish Standards Organization Committee S24 "Information and Documentation" proposed a general solution for the transport of MARC records using an XML container and suggested it be a supplement to the ISO 2709 standard (the standard which defines the record structure for MARC) [Andresen, 2003]. In discussion, it was agreed to develop this and make it a separate standard: work on developing this was due to start after an ISO/TC46/SC4 [2004] meeting in November 2004. The result of this would be based on the study of at least the Library of Congress's MARC21 XML Schema and the Open Archive Initiative (OAI) general XML container for MARC records. The agreement on such a standard would prevent many methodologies springing up and lead to ease of implementation of MARC XML by defining one methodology.

5.2.6 ONline Information eXchange (ONIX)

Since MARC does not well suit the needs of the booktrade which needs data not found in MARC such as cover design, synopsis, reviews and author biography [EDItEUR, n.d.], it has not been used as a bibliographic information standard for the book industry. Obtaining the data about each book from publishers to booksellers has been a challenge, complicated by the fact that each major industry company database has a different format preference for receiving the data. This lack of a standard made it difficult and time-consuming for publishers to format and

exchange their book information. ONIX is the XML-based international initiative originally for representing and communicating a variety of metadata elements about book industry product information in electronic form. And now ONIX has extended its coverage from books to serials and other media [Green, 2001]. From my research interview, I learnt that the Library of Congress Digital Library team is experimenting with the possibility of applying ONIX metadata in their bibliographic records as one of their future plans on XML technology.

ONIX stands for ONline Information eXchange, and is an approach of international collaboration including EDItEUR, Association of American Publishers (AAP), Book Industry Communication (BIC) and Book Industry Study Group (BISG). ONIX version 2.0 released on July 2001 defines both a list of data fields about a book and how to send that data in an "ONIX message". ONIX specifies over 200 data elements represented in XML format, each of which has a standard definition. Some of these data elements, such as ISBN, author name and title, are mandatory; others, such as book reviews and cover image, remain optional. While most data elements consist of text, for instance, contributor biography, many are multimedia files, such as images and audio files. Exchanging these optional fields, cover images, author photos and so forth, is particularly innovative.

Additionally, there are ongoing efforts of mapping from ONIX to MARC under collaboration between the Library of Congress Network Development and MARC Standards Office, British Library and OCLC [ONIX, n.d.].

Not only is the information held in the ONIX record and not found in MARC invaluable for book selection and acquisition, it also enables libraries to enrich their OPACs, and make them look more like the Internet bookselling sites that library users are becoming used to seeing. On the other hand, it is going to take some time for library automation systems to integrate the ONIX record structure, to the same extent as they have integrated MARC. Also, many millions of MARC records exist for items in which the book trade is not interested. So for that material, the MARC records will remain paramount for some time to come.

5.2.7 Open Digital Rights Language (ODRL)

Digital rights management is emerging as a crucial challenge for the content community in the digital age. There is a certain amount of rights management metadata in most metadata systems: the fifteenth element in Dublin Core is the Rights element; field 506 and 540 in MARC are about rights control as well, though rights management is not well-provided for in MARC. XML-based MPEG-21 includes elements to support identification and description of digital

resources, handling and usage of content, intellectual property management and protection.

Initiatives that could allow interoperability and support content to be made available in safe, open and trusted environments are needed [Iannella, 2001]. The Open Digital Rights Language (ODRL) [Iannella, 2002] is a proposed language and made available by W3C as a "Note" for the Digital Rights Management (DRM) community for the standardization of expressing rights information over content. The ODRL encoded by XML syntax, is being implemented, as it can express rights statements. An example is the OzAuthors online ebook store [OzAuthors, n.d.], which adopts XML for expressing syntax and ODRL as rights language. The goal of OzAuthors is to provide an easy way for authors to forward their content to the market place in an economic way with maximum royalties to content owners.

Although we have seen that licensing control exists in electronic journals controlled by publishers, electronic book rights management infrastructure seems more complex and is also of major concern due to the nature of the books and the fact that a book is of considerable value to its author in intellectual property terms [Snowhill, 2001].

5.3 Metadata Interoperability

The technology surrounding metadata is changing rapidly. Interoperability and long-term persistence are constant themes in the study of metadata issues, since a computer program does not recognize the structural metadata used to store a digital object in an independent repository. XML technology with the features of flexibility and open environment could provide a solution to this problem. There are a number of ongoing initiatives to define XML mechanisms for delivering metadata over the Internet. From the discussions in the early chapters, there is evidence indicating that XML is poised to become the standard transfer syntax for metadata on the Web. Below I discuss two XML-based metadata initiatives for interoperability. The W3C Resource Description Framework is an XML-based transfer syntax for exchanging complex metadata, and a mechanism for machine processing of metadata; the Metadata Encoding and Transmission Standard also is an XML-based framework for encoding all relevant types of metadata (descriptive, administrative and structural) used to describe digital library objects.

5.3.1 Resource Description Framework (RDF)

The Resource Description Framework, developed by the World Wide Web Consortium, is a mechanism for processing metadata. It provides interoperability between applications that exchange machine-understandable information on the Web [W3C RDF, 2004].

The history of metadata at the W3C began with the Platform for Internet Content Selection (PICS), which was originally designed to help parents and teachers to control what children may access on the Internet [Lassila, 1997].

RDF was a collaborative design effort by several W3C member companies. While RDF started as an extension of the PICS content description technology, it also drew upon the XML design as well as the technology proposals submitted by Microsoft (XMLDATA) and Netscape (XMLMCF). Other metadata communities, such as the Dublin Core (DC) and the Warwick Framework (WF), a container architecture for diverse sets of metadata (an approach from Digital Library Research group at Cornell University), had also influenced the RDF design [Berners-Lee and Swick, 1999].

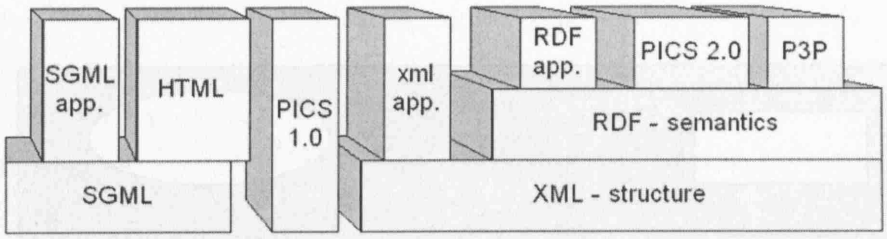

Figure 10: Data format architecture [Berners-Lee and the W3C Team, 1997]

Figure 10 gives an overview of some of the W3C data format specifications, and the relationships between them. RDF builds on the base of PICS and XML. XML replaces SGML and allows the expression of structure; RDF allows the expression of semantics.

RDF uses XML as a common syntax for the exchange and processing of metadata and supports the standard mechanisms for representing semantics that are settled in a simple, yet powerful data model. RDF develops a framework for the declaration of vocabularies to specify a set of vocabularies defined by a particular resource description community. Vocabularies are a set of properties, or metadata elements that can be reused, extended and refined to address application

or domain specific descriptive requirements. The mechanism, which was discussed in Section 2.3.3 of Chapter 2, is known as the XML Namespaces (W3C Rec 14 January 1999), a facility which is adopted in RDF. XML Namespaces support the concept of collaboration and reuse enabling resource description communities to keep and share the data element definitions among different parties [Miller, 1998].

The RDF data model is a syntax-neutral way of representing RDF expressions designed to impose structural constraints on the expression of models to support consistent encoding exchange and processing of metadata. Hence, it enables resource description communities to define their own semantics and provides for structural interoperability among applications. According to Lassila and Swick [1999], the basic data model consists of three object types: all things that can be described by RDF expressions are called resources; a property is a specific aspect, characteristic, attribute, or relation used to describe a resource; a specific resource together with a named property plus the value of that property for that resource is an RDF statement.

In the diagrammatic form in Figure 11, resources are always represented as ovals. Properties are always represented by arrows which point from the subject of a statement to the object of the statement.

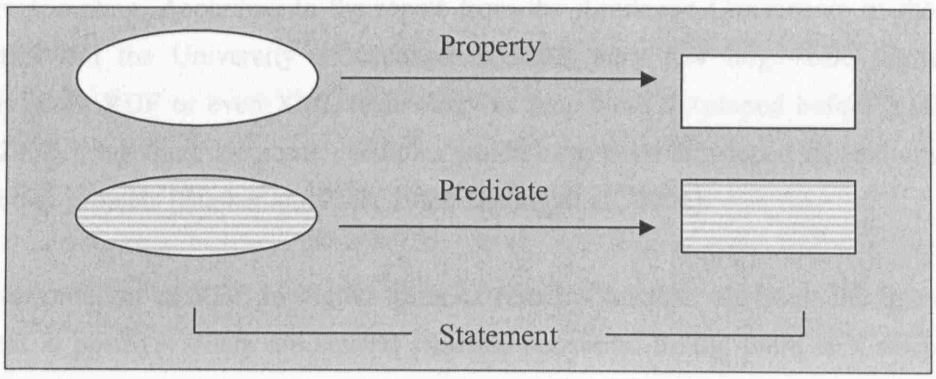

Figure 11: RDF basic data model

RDF can be used in a variety of application areas: firstly, in resource discovery, it provides better search engine capabilities; in cataloguing, it describes the content and content relationships available at a particular Website, page, or digital library. It uses intelligent software agents to facilitate knowledge sharing and exchange; in content rating, it offers a way of labelling resources, so that people (or computers) can filter information; in describing

collections of pages that represent a single logical "document", it describes the intellectual property rights of Web pages, and expresses the privacy preferences of a user as well as the privacy policies of a Website; finally, in digital signatures, it expresses information concerning what one is signing, what the significance of the signature is, the dates that the signature is valid and so on. RDF with digital signatures could be the key to building the "Web of Trust" for electronic commerce, collaboration, and many other applications [Lassila and Swick, 1999]. For example, in the digital library context, authenticity may become important, particularly in the science and technology fields, to contain the information as to who really did make a scientific discovery. RDF can facilitate that.

RDF is merely a way of representing data. The biggest challenge of RDF is to promote consistent deployment and create practical applications of RDF technology. There have been a number of commercial and research groups are developing RDF software and applications [W3C RDF, 2004]. Since interoperability is an important issue for the future Semantic Web, many of the RDF approaches have concentrated on integration with different technologies for representation and interchange of data on the Web. The potential benefit of this is that it will lead towards the integration of the heterogeneous information sources available in a digital library environment. For example, RDF integrates with the SOAP and WSDL (Web Services Description Language) in Web Services and knowledge management technology Topic Maps [Prud'hommeaux, 2001; Ogbuji, 2002]. The implementation of RDF in the digital library area is rather disappointing. According to the report from the Academic Consortium of the Big Ten Universities and the University of Chicago in 2002, very few large-scale digital library initiatives adopt RDF or even XML technology as they were developed before XML existed [UIUC, 2002a], but there are some examples which have been developed in medium or small digital library projects [Bunker and Zick, 1999; Habing et al., 2001].

The implementation of RDF in digital libraries remains unclear, although the initial overall assessment is positive. There are several possible concerns: firstly, there is a steep learning curve associated with RDF, and the value of RDF needs to be better support by richer technical tools; secondly, the algorithms for a working metadata system, such as DCQ, in RDF, and more complex modulated metadata structure in RDF remain unclear; finally, there is minimal guidance on encoding metadata system using RDF available in digital library research. Perhaps more collaborative projects between local or international institutions should be encouraged to further facilitate RDF applications.

5.3.2 Metadata Encoding and Transmission Standard (METS)

Since digital library projects are facing a challenge with the need to manage large and varied amounts of digital objects, the Metadata Encoding and Transmission Standard schema has been developed to provide a flexible mechanism for encoding and managing descriptive, administrative and structural metadata for digital projects. METS began in May 2001 from the Making of America II (MOA2) testbed project (a DLF project led by the University of California at Berkeley), which attempted to address the issues of an encoding format for descriptive, administrative and structural metadata for textual and image-based works [Library of Congress, 2001a]. METS addresses problems not on text markup, but on providing a structure for a digital collection or object by embedding and linking the parts that comprise the whole. This can be seen in the structure of METS documents shown in the following paragraph. From the document structure, it seems that the METS structure borrows from the concept of existing metadata systems such as TEI Header, and leverages the strength of XML technology, making it more powerful, so that it can manage the complex digital objects in a digital library environment.

METS is based on XML and is maintained in the Network Development and MARC Standards Office of the Library of Congress. A METS document is made up of seven major sections [Library of Congress, 2003c]:

- **METS Header** the METS Header contains minimal descriptive metadata about the METS documents itself such as creator, date of creation.
- **Descriptive Metadata** the Descriptive Metadata section may point to descriptive metadata externally to the METS document, for example, a MARC record in an OPAC or an EAD finding aid maintained on a WWW server; or contain internally embedded descriptive metadata; or both. Multiple instances of both external and internal descriptive metadata may be included in the Descriptive Metadata section.
- **Administrative Metadata** the Administrative Metadata section provides information regarding how the files are created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object, that is, master/ derivative file relationships and migration/transformation information. As with Descriptive Metadata, Administrative Metadata may be designed to work externally or internally.
- **File Section** the File Section contains one or more file groups which list all files comprising a single electronic version of the digital object.
- **Structural Map** the Structural Map is the main component of a METS document. It

defines a hierarchical structure for the digital library object, and links the elements of the structure to content files and metadata that pertain to each element. The element structure can be represented as follows:

```
<structMap>
  <div>
      ......
    <mptr> [external]
    <fptr>  [internal]
          ......
  </div>
</structMap>
```

Hierarchy can be represented within a single METS object (fptr) or through a series of related METS objects (mptr). <fptr> has child elements for sophisticated linking between the Structural Map and the File Section.

- **Structural Links** the Structural Links are designed to allow users to record the existence of hyperlinks between objects within the Structural Map.
- **Behaviour** A Behaviour section can be used to associate executable behaviours with content in the METS object, for example, page turns in a digital book, or indication of hierarchical or other arrangement in a collection of image files. A Behaviour section has an interface definition element that represents an abstract definition of the set of behaviours. The Behaviour section also contains the executable code that runs the behaviours defined by the interface definition.

METS is reviewed and endorsed by the Digital Library Federation and has become a digital library standard at the Library of Congress [Library of Congress, 2005b]. However, the advantages and limitations of METS would need to be recognized through further large-scale actual practices. From the research interview, I learnt that the Library of Congress has initiated the implementation of METS in their central digital repositories. More METS projects worldwide have been planned, are in progress, or have been fully implemented [Library of Congress, 2005a]. In Section 2.1.1.4 of Chapter 8, I will discuss further in detail the strengths, the advantages and challenges of METS in a digital library environment.

5.4 Metadata and the World Wide Web

The development of the World Wide Web has given metadata a more strategic role. I look at two different aspects of metadata, which have gained a certain amount of attention and have the potential to be used for resource discovery in Web-based systems.

5.4.1 Semantic Web

The Semantic Web is a vision of the creator of the World Wide Web, Tim Berners-Lee, with the idea of providing semantics for and facilitating the extraction of knowledge from the Web. The underlying concept of the Semantic Web is that the Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help [Berners-Lee, 1998].

One real power of the Semantic Web is that it is a plan for achieving a set of connected applications for data on the Web in such a way as to form a consistent logical Web of data. The fundamental layer in this infrastructure is the metadata associated with web resources. Metadata allows the Web to describe properties about some given contents. This means that the Semantic Web brings structure to the meaningful content of Web pages, creating an environment where software agents can readily carry out sophisticated tasks for users. This kind of application can be found in areas such as electronic commerce.

Another real power of the Semantic Web is that it can assist the evolution of human knowledge as a whole by joining together independent knowledge groups via URI. This structure would open up the knowledge and workings of human beings to meaningful analysis by software agents, providing a new class of tools which might bring in a new generation of Web-based knowledge-based systems. Knowledge portals in digital libraries could be one of the potential applications [Berners-Lee et al., 2001].

The Semantic Web is based on RDF, which integrates a variety of applications using XML for syntax and URIs for naming. Research in this subject is ongoing and active. This can be seen as starting from 2001 with the first conference on the topic of the Semantic Web: the International Semantic Web Working Symposium (SWWS) was held at Stanford University in the United States [SWWS, 2001]. There are a number of commercial products on the Semantic Web market. Some of the products try to lead the way to next generation semantically integrated

solutions by providing models that will be able to describe a complex domain of knowledge within an organization or environment [Network Inference, 2005]

In the European Union Technology Watch Briefing on XML, the Semantic Web is described as facilitating efficient resource discovery services, as well as reusing and repurposing across different platforms and software applications [Donnelly, 2003]. The difficulty of the Semantic Web is that it can only be reached when the XML family of specifications is in place and there is consensus on the content of the metadata elements, and so it is in the future rather than in the present.

5.4.2 Topic Maps

A topic map is a representation of information used to describe and navigate information objects. Several topic maps can provide topical structure information about the same information resources. Topic Maps provide access to information objects by means of different kinds of metadata, mostly but not exclusively what librarians would call subject data (including controlled and uncontrolled natural language, for example keywords) or classification data. They can also convey concepts like associations (relationships), occurrences (resources), or any other way that subject metadata can be organized. This type of metadata which I have just referred to as subject data is a structured view over a set of information resources that itself need not to be structured; that is, Topic Maps can be applied from "above" the information set, rather than from "inside" them; they are a superimposed view [TopicMaps. Org, 2001].

Topic Maps enable multiple alternative models of knowledge domains to coexist, and to work together. They help organize and retrieve online information in a way that can be mastered by information owners and information users, performing the same role as indexes play in books, and thesauri play in editorial consistency management. Topic maps are finding aids for the Web; they can be formatted as specific kinds of aids, such as ontologies, glossaries, thesauri and the like [Biezunski and Newcomb, 2001]. Examples of the data that could be included as topics in Topic Maps are endless. They can range from the language of content which could be identified by the name of the language expressed in English or as a language code such as the International Standard for Language Codes (ISO 639); or they could be topical and based on a standard classification scheme such as UDC or DDC, or any other faceted or non-faceted scheme [Bater, 2004]. More research has been undertaken on Topic Maps being represented in RDF data [van der Vlist, 2000; Lacher and Decker, 2001; Bater, 2004].

In an academic digital environment, Topic Maps can be strengthened by the link between the

academic fields of knowledge organization with its principles and methods, and knowledge management with its application-oriented standards [Sigel, 2000]. The topic maps mechanism is s expected to play a key role in handling semantics computationally in the next generation of ontology- and agent-based knowledge services.

In an academic digital environment, Topic Maps can be strengthened by the link between the academic fields of knowledge organization with its principles and methods, and knowledge management with its application-oriented standards [Sigel, 2000]. The topic maps mechanism is expected to play a key role in handling semantics computationally in the next generation of ontology- and agent-based knowledge services.

Topic Maps are defined in an international standard (ISO/IEC 13250:2000), and its technology is progressing via the production of XML-based Topic Map serialization syntax (XTM), written by the members of the TopicMaps.Org Authoring Group. This specification provides a model and grammar for representing the structure of information resources used to define topics, and the associations (relationships) between topics [TopicMaps. Org, 2001].

The Topic Maps framework is still work in progress. Tools and consulting companies are emerging on the market. I observed a seminar on this topic in the Library of Congress held by a commercial company when conducting my research interview there. In general, more pilot projects would help people to know and use Topic Maps better, and that could improve future knowledge information systems in the area, such as digital libraries. The Semantic Web and the Topic Map are not yet used very much, but the two efforts have potential being key technologies for knowledge management in digital libraries.

5.5 Prospects for XML-Based Metadata Initiatives

The world of the digital library is a fast changing world. However, metadata depends for its effectiveness on the avoidance of obsolescence. XML has been embraced by organizations which have already developed their own metadata standards. In many cases, an XML version of existing initiatives has been developed. Dublin Core does not require XML, but in RDF, Dublin Core can be embedded in XML. The future of Dublin Core was promising when it was developed and adopted by OCLC which has the objective of using it as a standard for the cataloguing of the universe of bibliographic materials in digital form, for example in the CORC project, but its future is now even more certain, since it has been adopted as an international

standard. The main problem is the lack of semantic equivalence between, for example, different subsets of Dublin Core (qualified Dublin Core) or between Dublin Core elements and those of other metadata sets.

TEI Header is supported by a wide range of institutions as we have seen above. EAD has the support of the International Council of Archives and the LC NDMSO. CIMI was an important organization in museum documentation and had the cooperation in the development of the SPECTRUM standard of its equivalent organization in the United Kingdom, the MDA. MARC is supported by the LC NDMSO and ONIX by BIC, which is an organization leading the publishing industry into standardization and automation to the same extent as libraries have computerized; ONIX also has the support of the Library of Congress. Although ODRL is only a "Note" from W3C, it still has the potential to be adopted by the communities.

These XML initiatives are very much supported by leading institutions which, I believe, ensures they will survive into the future and will not remain just as projects of researchers, and thus have the potential of being important in digital library development. At the same time, XML is becoming a popular medium for the exchange of data in general on the Internet, outside the library and information fields. I believe that this will help foster the growth of XML-based initiatives in digital libraries. The evolving metadata standards require an infrastructure to facilitate modular interoperability among diverse, application-specific metadata systems. But only when there has been plenty of experience through real-world initiatives and the issues have been discussed at greater length than is the case at present, will it become clearer how the XML-based metadata initiatives will function in digital library development.

Part II

Part two of the thesis consists of Chapters 6-9, which is a case study of three digital library initiatives, and Chapter 10, which is the conclusion. Chapters 6-9 comprise research issues including the rationale for digitization, realizing the digital libraries and maintaining the digital libraries, and they are organized in a structured pattern: interviews, analysis and conclusion. Throughout Part Two, Perseus refers to the Perseus Digital Library, Michigan refers to the Michigan Digital Library Service, and LC is used for the Library of Congress. Although the National Digital Library Program at the Library of Congress is officially finished, as indicated in Section 3.3 of this Chapter, digitizing work in the Library of Congress is ongoing. Many new projects and new collections which may share the same resources and technologies with the National Digital Library Program have been added to the Library of Congress National Digital Library, and integrated into the Library's production system.

Chapter 6

Case studies

In this Chapter, I introduce the three digital libraries; describe the rationale behind the research interviews and the methodologies adopted in this part of the research. In addition, I describe the background to the case studies which are the University of Michigan Digital Library Service (DLS), Perseus Digital Library (PDL) at Tufts University, and the Library of Congress National Digital Library Program (NDLP).

6.1 Introduction

The library community is actively taking advantage of the enormous potential use of information and communication technologies to enable their collections to be available online. They are facing the challenges of the rapidly changing and growing digital environment, and developing practices to best support its use and management over time.

The community recognizes that these issues and challenges are not only technical in nature, but affect the organizational situation, user demand, and intellectual and economic aspects. To support this, they are building digital repositories involving community groups as collaborators. They are exploring innovative approaches on how to manage physical, digitized and born-digital objects integrated as a whole into their collections, and on how to preserve the value of digital resources in a rich interactive digital environment. They are developing tools for providing flexible and powerful access to the repository with the least technical barrier. They are exploiting business opportunities and are being designed to be economically self-sustaining. They are highlighting surrounding legal and policy framework issues associated with making information available. They are making strategies for institutional change and integrating the results of these strategies into the current workflow to overcome existing technological, organizational and legal impediments [Greenstein and Thorin, 2002].

6.2 Methodology

In order to further my research and facilitate my understanding of the real world digital library development, I selected three digital libraries as case studies and planned to involve both quantitative and qualitative research methodologies, including in-depth research interviews, Web usage statistics and user surveys.

Firstly, an interview questionnaire was carefully designed aiming to explore the research areas including collection management, conversion methods and processes, technical infrastructure to deliver digital library contents, organization strategy and funding and, in particular, the impact of the role of XML in every possible aspect of the development of the digital library. The research interviews provided the discussion subjects which formed Chapter 7, 8 and 9.

Before I conducted the interviews, the interview questionnaire was sent in advance to Caroline Arms at the Library of Congress, John Price-Wilkin at the University of Michigan Library and Gregory Crane at the Perseus Digital Library. The research interviews took place during 12–19 September 2002 with the support of an award from the Graduate School Research Projects Fund at University College London. I did the first interview with the Perseus Digital Library at Tufts University with the principal research members of the group including Gregory Crane, Clifford Wulfman, Anne Mahoney and Thomas Milbank. In addition, I interviewed via emails David Smith, former Head Programmer at Perseus, who at that time was doing a research degree at the Johns Hopkins University. I did the second interview with two of the staff members in the

University of Michigan Library, the Associate Librarian, Christina Powell and Image Collection Coordinator, John Weise. The Library of Congress National Digital Library Program was my last interview. During this visit I was fortunate in being able to interview as many staff as I did, although there was a three-day XML-subject workshop held in the Library, which many staff were appointed to attend. I interviewed Head of the Network Development and MARC Standards Office, Sally McCallum; Technical Coordinators at the Office of Strategic Initiatives, Caroline Arms and Martha Anderson; the Library of Congress Digital Audio-Video Preservation Prototyping Project Coordinator, Carl Fleischhauer; Project Coordinator for the digitization of manuscript collections, Laura Graham; and Project Coordinator for the digitization of books and general collections, Steven McCollum; computer specialists at the Information Technology Service, David Woodward and Mary Ambrioso.

I sent some follow-up emails after returning from the visit to request supplementary information. The interviews were then transcribed into text and organized systematically into different research subjects. The interview questionnaire is reproduced in Appendix I, and the transcripts are saved on a CD-ROM in Appendix III of the thesis. The CD-ROM is attached to the back of the thesis. The interviews are the main source of information in the following chapters except where indicated otherwise.

Secondly, I collected the HTTP Web-usage data on the three digital libraries through research interviews and generated statistical graphs and made comparisons between the three. The analysis and conclusion of the statistical results are covered in Chapter 9, Section 1.1.3 Statistics.

Thirdly, I designed a user survey questionnaire aiming to investigate the interface design, response time and the tools and help features provided on the three digital library websites. At the same time, I requested permissions from the three digital libraries to cooperate with me in the user surveys. The final questionnaire is reproduced as in Appendix II.

After several follow-up emails, Michigan, however, replied that they were not able to cooperate with my user survey due to political and technological problems. I also received a negative reply from the Library of Congress on the grounds that user questionnaires were not consistent with LC policy. Perseus never replied despite three requests, and so I had no way of knowing their intention on this matter. Since two of my three case study digital libraries were unwilling to cooperate with the user surveys, and the results of any Perseus user survey would have been unlikely to illustrate the success of XML without being able to make comparisons with the user

survey results of other digital libraries even if Perseus had agreed to cooperate eventually, it would not have been very profitable.

6.3 Case Studies

In Section 1.3 of Chapter 3, I discussed how there has been richer investment in digital library initiatives in the United States than in the United Kingdom; research institutions and government agencies in the United States have a long record of cooperation for joint efforts in the digital library arena. For example, the DLI-funded digital library project at Stanford University is represented in a large gathering of more than seventy industrial partners of Stanford University's database group [Stanford Digital Library Project, 1997]. One of my case studies, the Library of Congress NDLP is also an example. Furthermore, I found more satisfied large-scale XML and SGML digital library approaches in the United States than in the United Kingdom. With strong financial and technology support, it has been easier for digital libraries in the United States to move from research to practice and from prototypes to operational systems and services. It is mainly the promise of rapid and in-depth access to information that has been fueling the need for increased investment in technology; meanwhile, commercial technology enterprises have been actively contributing actively their technologies to this research area. The LC NDLP is a case, which I will discuss in this Chapter.

There has been significant development in digital library initiatives in the past decade, and I found many notable examples of best practice. My case studies focused on the United States instead of on the United Kingdom for reasons I discussed in the previous paragraph. The three digital libraries were selected because they used SGML and XML and were exemplars in the use of this technology to manage large amounts of data. These quantities were not found in digital libraries in the United Kingdom or in other digital libraries in the United States. Their practices provided excellent case studies for this research. Also, the three initiatives were selected because the three represented distinct characteristics of digital library development. The University of Michigan Digital Library Services is a campus-wide research-based academic digital library programme, and is fully integrated into its larger organization, the University Library, while the Perseus Digital Library project based at Tufts University is entirely a research testbed to explore the technology in developing a digital library without the strong institutional support that Michigan has. The Library of Congress plays a leading role as a national library. Its strategic planning has been regarded as providing the best guidelines in developing a digital

library programme, and it has a strong influence on the library and information community in general.

6.3.1 University of Michigan Digital Library Services (DLS)

Introduction and background

The University of Michigan Digital Library Services is one of the oldest digital library programmes in the United States. It was launched in 1996. The catalyst was the campus-wide faculty 1991 Information Symposium discussion on how the library needed to adopt new technology and transform itself into a networked environment [School of Information and Library Studies, University of Michigan, 1991]. A report was prepared for campus administration with three recommendations on the issues of the intensive analysis of distributed computing operations and the economics of campus information provision. The recommendations were that the campus should attempt to:

- Bring together library and technology expertise;
- Develop visible projects;
- Create an "information community".

Eventually, the Digital Library was sponsored jointly by three principal information organizations on campus: the Information Technology Division, the School of Information and the University Library. The funding supported the creation of a Digital Library Programme Director position in charge of a separate department within the University Library. This arrangement proved positively later that the Digital Library could thereby have more room to experiment with new initiatives without affecting traditional library management, but supporting each other in every possible library principle [Price-Wilkin, 1999].

The Digital Library defined its scope in both technical and organizational criteria that would focus on developing a network information environment in support of the academic community with coherence and integrity. It addressed not only the technical issues in developing an information environment, but the significant behavioural and institutional issues as well. Consequently, attention has been paid to involving the campus in articulating information needs, exploring the financial, legal and administrative questions surrounding a distributed information environment and encouraging the University's investment in information and information technology.

Meanwhile, a project-based programme strategy was embarked upon. The early activities focused on issues of durable formats for digital content building and developing architecture for digital collections. The content-building projects were based on work in the late 1980s where the University Library began some preliminary efforts to build digital library components [Lougee, 1998]. It firstly started with non-bibliographic information sources such as GIS data, and then a formal access system for text encoded in SGML.

Partnership and collaboration

Michigan's broad and campus-wide partnership has played a crucial role in promoting the concept of the digital library, and seeking to reach the maximum potential of the digital library [Lougee, 1998]. This partnership and collaboration have promoted its goal of achieving a comprehensive, coherent networked-information environment and, moving beyond, extending partnership relationships off-campus including to the publishing and library community. A number of significant initiatives have been undertaken based on this concept. The following are some of the efforts.

The University LIcensing Program (TULIP), which started in early 1991 and concluded at the end of 1995, was a collaborative project between Elsevier Science and nine leading universities in the United States. The goal of the project was jointly to test systems for networked delivery to, and use of journals at, the user's desktop [Willis, 1995]. TULIP contributed substantial parts of the infrastructure in the development of the University of Michigan Digital Library. For example, the creation of the search engine FTL is now refined and used in JSTOR (discussed below). Michigan was the first site to implement the forty-three journals in materials science offered through TULIP, and was also the first to move the service to the Web environment. More than this, the TULIP effort contributed to the partnership in bringing the Library to work effectively together with another campus entity, the School of Information [Price-Wilkin, 1999]. At this point, Michigan was awarded funding by the Digital Library Initiative Phase 1 (1994-1998).

Michigan's participation in the TULIP project was the reason why it was selected by the Mellon Foundation in the Summer of 1994 to receive a grant to extend the work Michigan had done as part of the TULIP project known as JSTOR. Its original purpose was to investigate whether it was possible to increase access to older scholarly journals (provided by Elsevier) by converting them to digital media while simultaneously ensuring their preservation and saving library shelf space [Guthrie and Lougee, 1997]. By 2005, JSTOR had become an independent not-for-profit organization growing in size and scope, and the database itself is housed on a server at Michigan. JSTOR is available to academic institutions via site licences. The JSTOR effort

provided Michigan digital library staff with best practice policies about archiving and retrospective conversion.

The Michigan administrator strongly believed that the long-term success of digital libraries would be heavily relied upon for solving payment mechanisms in the Michigan system architecture [Atkins, 1996]. TULIP and JSTOR address a broad set of issues including technical and user behavioural interest, but none of them assesses issues of economics and pricing for electronic journal publishing on the Web. In 1996, with further cooperation from Elsevier, Michigan launched Pricing Electronic Access to Knowledge (PEAK), which was a large-scale collaborative effort between the University's Programme for Research on the Information Economy (PRIE) and the Michigan digital library programme to explore pricing models for electronic journals. Through project PEAK, Michigan staff could continue Internet distribution research, and work together with economists, psychologists and Human Computer Interface (HCI) experts to investigate database mechanisms for authentication and for subscription/ purchase control information [Bonn et al., 1999].

The Humanities Text Initiative (HTI) was launched in 1994 with joint sponsorship from the University Library, the School of Information and the University Press. HTI's initial work was based on the University Library's efforts in 1988 known as UMLibText, an Internet-based textual analysis capability service. HTI was designed as an umbrella organization for the creation and maintenance of the online text, and as a mechanism for strengthening the University's capabilities in the area of electronic text. HTI creates in-house electronic texts and also delivers SGML-encoded pages created by outside publishers or electronic text vendors, such as Grolier, Oxford University Press or Chadwyck-Healey. In 2005, there were more than five million pages of encoded text created by HTI, which is one of the largest and richest collections in SGML available on the Internet. With its experience in electronic text, HTI started to provide a service to assist other academic institutions in the development of electronic text in SGML through the SGML Server Program (SSP) [Powell and Kerr, 1997], which later was replaced by DLXS, a broader range of production service in Michigan.

The Making of America (MOA), Part 1 represented a major collaborative endeavour between the University of Michigan and Cornell University. The MOA1 effort concentrated on preserving and making accessible through digital technology a significant body of primary sources related to American social history from 1850 until 1877. MOA2 was launched in 1996 and was expanded to a consortial initiative among Digital Library Federation members. More emphasis was placed on the issues of interoperability, scalability and digital preservation [MOA2, 1998].

Production service

The project-based initiatives were greatly helped by the evolution of expertise on the campus environment at Michigan. The administrators believed that there was a need to integrate fully a range of project activities on campus into one production service, since many of the projects had been ongoing, producing several arrays of digital elements. In 1996, the University Library, Media Union, Information Technology Division and the School of Information jointly supported the creation of the Digital Library Production Service (DLPS), managing to sustain and develop digital library collections. DLPS cooperates with the four units to ensure it is meeting the goals set out by them.

DLPS was established with the intention of providing long-term support to the growing array of production digital library operations, but also to take lessons learned from previous efforts to develop creative thinking in designing, creating and maintaining the mechanisms needed to deliver library information via networked mechanisms. To this end, DLPS has built a core infrastructure, including various object classes and interchange formats that could be supported across media.

Having gained vast experience from these highly focused projects, Michigan experimented with new technologies and gained competencies in many areas. The JSTOR project and MOA effort provided a range of experiences associated with large-scale digital reformatting, while TULIP and PEAK projects provided data management experience as well as experience in the pricing of electronic resources. In late 1999, DLPS strengthened its digital library production operations, and created the University of Michigan Digital Library eXtension Service (DLXS) to replace the SSP. DLXS offers educational and non-profit institutions a powerful search engine and an array of class-based middleware for mounting numerous and heterogeneous digital objects, which have for years served as the key tools for digital library services and resources at Michigan. DLPS sells memberships to institutions enabling them to receive its tools and technical support, so that the institutions can fully develop their digital library collections. According to DLXS's latest updated website, this membership group had expanded to 29 institutions worldwide by the end of 2003 [DLXS, 2003b].

More integration

Later, the Digital Library was renamed the Digital Library Services (DLS) to reflect its broader mission in serving the University Library and the information technology community. DLS expanded into five units: Library Systems Office, responsible for managing the Library's online catalogue and related tools; Desktop Support Services, responsible for managing the computing infrastructure of the Library; Digital Library Production Service, responsible for managing and

developing digital collections; Core Services, responsible for Unix/Linux system administration, as well as broad system integration; Web Services, responsible for managing the Library's public and staff web space. This arrangement was with the expectation that Michigan could provide an overall service in dealing with issues such as access to electronic resources, networked information, digital collections and information policies.

According to Price-Wilkin [1999], the Michigan sponsors believed that the progress of digital libraries would depend on the creation of production organizations because of the need for continuity and reuse of resources in a coordinated and planned way. The administrators also recognized that when developing a digital library, the production organizations must be fully integrated into the campus's academic mission as well as into the University Library's mission and function. Recognition of these facts has contributed largely to today's University of Michigan Digital Library, and successfully boosted it into a national and internationally known leading digital library in respect of its initiative and its innovative practices.

6.3.2 Perseus Digital Library (PDL)

Introduction

The Perseus Digital Library project has been under continuous development since the spring of 1987, but could even be traced back to 1982 when it was merely a concept, when Professor Crane was doing research work in full-text retrieval in Greek and other languages at the Harvard Graduate School. The project is still directed by Professor Crane who leads a small team of classicists, and after a number of years at Harvard, it is now hosted at Tufts University. The goal of Perseus reflects the ambition to revolutionize the study of the Humanities by expanding the ways in which literature, history, art and archaeology can be examined efficiently. As Professor Crane said to this author, the work of the Perseus project is not only to help traditional scholars conduct their research more effectively, but more importantly, to help scholars in the Humanities to use the technology to redefine the relationship between their work and the broader intellectual community.

Greek Perseus

The Perseus' initial work started with Greek Perseus, using this domain as a testbed to become a digital library for the Humanities. There were a few reasons why Greek was selected: firstly, it was Professor Crane's subject, but also he thought there was a finite body of materials in classics. The field was interdisciplinary, and it was a difficult area that presented the challenge of being in Greek [Harris, 1999], with its use of a different script.

The Perseus team believed that language stood at the heart of any database of cultural materials because the two are so inextricably linked together, and the system thus should help its users work with the languages [Crane, 2000c]. This idea later drove the work on digitizing source material, developing tools for searching and analyzing pre-modern languages, and creating an editorial process for electronic publication. Greek Perseus completed works of thirty-one authors in classical Greek texts, many of them accompanied with notes. The collection also consists of twenty five thousand images covering architecture, sculpture, coins, vases and sites, making Perseus the largest photographic database of ancient Greece ever published. In addition, an online atlas became available at this stage.

Greek Perseus created electronic resources as a core database repository in SGML, and with vision converted it into a Macintosh-based Hypercard form for distribution, which was a tool for authoring a media-rich interactive solution, the only one available at the time. According to Professor Crane, Greek Perseus only used the Hypercard environment as a delivery vehicle because creating a new version of Perseus Hypercard would have required a great effort in porting the data from a variety of more specialized formats, for example SGML texts, and it proved efficient when changing from Hypercard into a World Wide Web environment in 1996 [Crane, 1998].

Roman Perseus

In the summer of 1996, with initial support from the Teaching with Technology Programme at the National Endowment for the Humanities (NEH) and source materials from Tufts University, Perseus began the work of establishing a reasonably substantial, well-integrated and open-ended Roman Perseus. It built on the success of the tools and resources developed for Greek Perseus, but with additional art and archaeology materials as well as new collections of Latin texts and tools [Crane, 2000b].

Perseus created new ways of working with classical materials in Roman Perseus, in which the visualization technique was the one with the most potential. More techniques became available such as better control over the display of text, especially the ability to use different Greek fonts, automatic hyperlink to places and other texts to help the users. As the support with funds was limited, Roman Perseus could not provide the same coverage for Latin as for Greek materials, but instead created a framework that could evolve and mature over time. Perseus created a database on the Roman world roughly one third as large as that on Greece.

Greco-Roman Perseus

The existing Greco-Roman digital library is a large body, making up a thematically coherent

testbed which includes valuable primary source documents, objects and sites studied in the Humanities. It covers five million words in Greek and Latin linked to accompanying English translations, three major Greek and Latin dictionaries, three grammars, a growing number of commentaries, thousands of descriptive catalogue entries for sites and museum objects, hundreds of site plans, four thousand toponyms that can be plotted on more than one thousand maps, thirty thousand colour images, most of which were taken for Perseus based on consistent standards, encyclopedia articles, essays, books and other resources.

By the time of the research interviews, Perseus had reached an established audience in academia, schools, libraries, and with the general public. With the possible technical limitation in the network as Perseus consists of large amounts of non-text materials, the release of the platform independent Perseus version 2.0 with four CD-ROMs containing the complete database by Yale University Press in 2000 supported beneficially the teaching and study of the ancient world. Some specialists would argue that CD-ROM as a medium is inappropriate in the 21st century. However, the Yale University Press was no doubt aiming the product at schools where Internet access can be a problem which can be avoided by using CD-ROM, especially with multimedia, given that teachers need fast access to learning materials when standing in front of the classroom. For example, the Graduate Institute of Education at Sun Yat-sen University in Taiwan uses Perseus on CD-ROM as a teaching medium [Yang, 2001].

Funding

Professor Crane defined Perseus as a research testbed, and so it has not integrated its mission or function with University developments. Therefore, Perseus does not receive strong support from divisions at the University or University Library, and raising external funding to expand its holdings has always posed the most challenging aspect for Perseus. Perseus has received support from a broad range of foundations, institutions, private donations and industries.

A big boost to Perseus' credibility occurred in 1999, when NSF/DARPA/NASA awarded Perseus a Digital Library Initiative Phase 2 grant for 2.8 million dollars. The grant offered to the Perseus project the opportunity to refine and maintain extensively the original work on Greek and Rome, and to conduct in-depth research on human and technical issues in digital libraries such as digital library evaluation. Furthermore, this grant allows Perseus to expand its on-going relationships with project partners [Crane et al., 1998].

Evolution

From the very beginning, the Perseus Digital Library was designed as a "mature" digital library that covers a wide range of subjects in the Humanities. It explores and develops tools to support

complex searching and information retrieval, and evaluates the impact of digital libraries on teaching and learning. To reach these goals, Perseus' effort has been working in four directions: large volumes of an array of media of primary and secondary resources; self-directed access and use; open environment approach; teaching and learning.

Firstly, there is the direction of adding large volumes of resources. Perseus is fundamentally a collaborative enterprise that has been working actively together with many scholars in the Humanities at a variety of institutions, hoping to develop closer connections with similar projects. The Museum of Fine Arts in Boston is one of the major collaborators. Through cooperation with the Museum, Perseus made a collection of images of objects including sarcophagi, bronze statuettes, funerary reliefs and many others, and created the Perseus Sculpture Catalog and the Perseus Coin Catalog.

Perseus built an English Renaissance collection, which was initiated in 1996 with work on the early modern period, an electronic edition of the complete works of Christopher Marlowe. Perseus contracted with the Modern Languages Association to create an electronic format on the New Variorum Shakespeare (NVS) series, the premiere series of scholarly editions on the plays of Shakespeare as a major Shakespeare resource for its English Renaissance collection. Three university libraries also contributed their holdings to Perseus' Shakespeare database: the Furness Shakespeare Library at the University of Pennsylvania, Brandeis University and special collections at Tufts University.

A major challenge came from the collaboration with Tufts University Archives on the Edwin Bolles Archive which possessed an enormous wealth of information about historical topography of London and its environs [Crane et al., 2001]. The London project focused on multimedia materials to document the physical context including visual materials such as still-images, sound, video, emerging tools for virtual tours and 3D representations of objects. In quantity, this testbed includes a corpus of more than ten million words, twice as many as the words in the Greco-Roman Perseus.

Additionally, Perseus also expanded significantly in developing electronic tools for the history of science through a joint project called Archimedes, with a large research group at the Max Planck Institute for the History of Science in Berlin.

Secondly, there is the direction of self-directed access and use. When interviewed, the Perseus team believed that digital libraries must help readers find and analyze the complexity of materials by automatically generating links between different electronic objects, establishing

consistency of document structures and references. Therefore, digital libraries should be able to provide tools for representing, accessing and visualizing these complex objects. To fulfil this goal, Perseus has developed and integrated a set of reference works and tools, which probably could be regarded as one of the most distinguished achievements in Perseus efforts.

In Perseus, electronic resources interact dynamically to allow for flexible access that supports different kinds of use. Citation documents or searched results are not just destinations but starting points; in other words, they are bi-directionally linked. Perseus started as an interface to the *Thesaurus Linguae Graecae (TLG)*, from which most of the Perseus Greek texts derived, and many of them were accompanied by English translations. The morphology parser for analyzing Greek texts is at the centre of the Perseus project. Every wordform in the Perseus corpus contains all its possible morphological description, and through this morphological parser it is linked to the lemmata or dictionary entry forms in *Liddell-Scott-Jones (LSJ)*, one of the most authoritative dictionaries of ancient Greek. All the citations in the dictionary are in turn linked back to the Perseus corpus. The English equivalent of this would be a digital corpus in which any wordform is linked directly to its lemma in the *Oxford English Dictionary,* and quotations for the lemma are linked back to the full texts from which they are taken. The parser was written originally to parse classical Greek, and it has been extended accordingly to Latin, Italian and German in Perseus. Other tools include a hypertextual encyclopedia of major historical and mythological figures, places and terms. An extensive atlas has been developed which includes schematic, satellite and topographical maps of ancient and present-day Greece, on which more than sixteen hundred sites can be plotted. These tools allow users to zoom in and out of maps, and explore a variety of spatial themes across space and time.

More tools have been developed in the Bolles London project. This project exposed the challenge not on how to reconstruct 18th century London but rather on how to represent the information manageably. To help readers grasp the complex temporal-spatial interactions, Perseus applied the Geographic Information System (GIS) as a tool to geo-reference historical maps, the *Getty Thesaurus of Geographic Names (TGN)* to locate places in London or in a broader geographic focus and the *Dictionary of National Biography (DNB)* to disambiguate personal name references. This is helpful when applied to general document analysis. This ideal for Perseus was to incorporate GIS and virtual reality technologies in a set of tools to improve visualization of the numerous temporal and spatial interconnections between the materials of London of that period. Furthermore, Perseus developed an automatically generated "Timeline", which acts as a pointer to the content and focus of documents, as well as providing browsing aids that would enable the user to go into greater levels of detail or go from the timelines to the relevant sections of the document [Crane, 2000a].

Thirdly, there is the direction of the open environment approach. It was the original idea that Perseus was built with the ability to scale this research testbed to full operational status, with ever-larger collections. And this would depend upon workable standards and interoperability. One of the greatest challenges for Perseus has been the need to manage documents with widely varying encodings and markup practices. Perseus has been benefiting greatly from the generality and abstraction of structured markup, SGML, and now moves to the new open initiative XML. Firstly SGML and later the XML served Perseus text and documents and conformed to the TEI, which is another international effort.

Meanwhile, Perseus is evolving in an open source environment. Perseus uses Linux Sharp primarily with the back end database management system PostgreSQL for a production Web server that delivers over two million pages a week.

Fourthly, there is the direction of teaching and learning. The variety of approaches in Perseus has made Perseus an interactive teaching tool. The Perseus user community has been an ill-defined group because it has a broad spectrum of users. One example for this might be a secondary school student taking a Latin course, who would find the text-and-dictionary environment useful.

To support teaching and learning, the Perseus team has been running a series of seminars for teachers as far as they could if support is available, for example Perseus provides the Stoa publishing group with the technical support and the ability to produce the electronic publication of materials. This programme, known as the Perseus Project Publication Model project, was funded by the Fund for the Improvement of Post-Secondary Education (FIPSE). This was a three-year (1997-2000) project that aimed to leverage the Perseus resources and World Wide Web infrastructure to develop and evaluate publication models in classics. In a broad sense, Perseus was helping Stoa to create new electronic publications that not only serve the traditional academic audience, but also exploit the technology to bridge the gap between academic publications and the wider audience [Marchionini et al., 2000].

More collaboration, more challenges

The Perseus team believes that, after fifteen years of work on these projects, they have provided models for how the best practices from traditional libraries and editorial efforts can evolve and adapt to meet radically changing circumstances. Thus, Perseus will continue to study how to respond to more challenges and possibilities of new digital media, if federated digital libraries are to realize their potential.

6.3.3 Library of Congress National Digital Library Program (NDLP)

Introduction

The Library of Congress collects materials from all over the globe. For many areas of the world, its collections are the finest and most comprehensive research collections outside their country of origin. The Library manages the largest and most varied archival collection of American creativity. The Library makes its collections available to the nation's schools, libraries and life-long learners through public reading rooms, inter-library loan, the Internet and copyright-compatible copying. The Library's mission is summarized as "to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations" [Billington, 1995a].

The National Digital Library Program (NDLP) is an initiative of the Library of Congress. It was a five-year programme, beginning in 1995. The group assembled a digital library of reproductions of distinctive and historical Americana primary source materials to support the study of the history and culture of the United States. These materials include photographs, manuscripts, rare books, maps, recorded sound and moving pictures, selected for digital conversion from the Library's vast holdings of print and non-print materials.

Apart from NDLP, in the Library, there are several projects that address the challenges of providing direct access for users to the Library's resources and services as well. The THOMAS service supports public access to the full text of all House and Senate bills of the past sessions of Congress and the complete text of the Congressional Record. The Copyright Office Electronic Registration, Recordation, and Deposit System (CORDS) developed in the Copyright Office, provides an effective source to accept new publications in digital form for copyright registration and deposit, which it was believed would also be of benefit to the growing NDLP collections. Although they are separate projects, these different projects shared the same tools and technological framework, as they must all be integrated into the Library's production system [Arms, C. R., 1996].

In 1998, the staff of NDLP identified ten challenges that should be met to build a large and effective digital library for the 21st century, expecting that through the sharing of creative and innovative ideas, it would be possible to formulate policies on these issues. These ten challenges were grouped into five broad categories, building the resource, interoperability, intellectual property, effective access and finally sustaining the resource [Library of Congress, 1998b]. The first challenge concerned developing improved technology for digitizing analogue materials which involved not only improving the hardware, but also the development of mechanisms for

making textual and non-textual material more interoperable, and one of these mechanisms involved using the XML or SGML. The second concerned designing search and retrieval tools that compensate for abbreviated or incomplete cataloguing or descriptive information. The greater availability of XML compatible data has the potential to assist in this area. The third, designing tools that facilitate the enhancement of cataloguing or descriptive information, by incorporating the contributions of users, will benefit from XML since there are many tools becoming available that will convert the contributions of users into the XML format. The fourth, a contribution to interoperability, involved establishing protocols and standards. XML provides a complete open environment which can facilitate the interoperability task. For intellectual property and rights management, the fifth challenge, addresses legal concerns associated with access, copying, and the dissemination of physical and digital materials. For this, XML-based tools such as DOI can be deployed. The use of XML also facilitates effective access such as integrating digital resources metadata and physical resources metadata into one catalogue for cross-material searching, which is the sixth challenge.

XML also helped the seventh challenge, which was to develop approaches that can present heterogeneous resources in a coherent way, since XML is a tool that encourages standardization. Because of the universal readability of XML through Web browsers, it would make the National Digital Library (if based on XML) available and useful to different communities of users and for different purposes which was the eighth challenge. One of the aims of XML is the same as the ninth challenge formulated by the Library of Congress: to provide more efficient and more flexible tools for transforming digital content to suit the needs of end-users. The tenth challenge which came under the heading 'Sustaining the resource' was to develop economic models for the support of the National Digital Library; here the universal availability of XML, its likely future survival and the large ranges of tools which are likely to be developed to use with it, mean that XML is a good choice to support the sustainability of a digital library.

Background and architecture

NDLP builds on the experience of two earlier pilot projects, the Optical Disk Pilot project and the American Memory project. The Prints and Photographs Division (P&P) started as an experiment in the mid-1980s with the Optical Disk Pilot Programme (1982-1987) for making electronic reference copies for its fragile negatives and transparencies. NDLP provided the infrastructure to make their treasures more accessible outside the reading room, as well as helping NDLP to further its mission. The American Memory pilot ran from 1990-1995 and emphasized digitizing the nation's memory, and also was consistent with the Library's archival practice [Fleischhauer, 1995]. The pilot established technical procedures, tackled intellectual property issues, demonstrated options for distribution, identified the audiences for digital

collections through an end-user evaluation, and began institutionalizing digitization activity at the Library of Congress.

NDLP created a multimedia dimension, such as written materials including books, and other printed texts, manuscripts, and sheet music, maps, motion pictures, photos and prints, and sound recordings. Materials were selected through internal divisions that were unique to the Library of Congress or rarely held elsewhere. In the pilot stage, in order to explore the problems of capture, storage and representation, these materials of various types were converted to both CD-ROM and videodisc formats, since the World Wide Web had not yet provided a network-based mechanism for accessing multimedia materials [Arms, C. R., 1996].

Immediately after the end of the American Memory Pilot, the increase in the usage of the World Wide Web ensured that American Memory was accessed extensively by the public through the Library of Congress's Website. The success of the pilot helped the Library to raise private funds and win congressional support to continue and expand the effort [Fleischhauer, 1995]. NDLP received funds of around sixty million dollars over a five-year period (1995-2000) from the Congress and private sectors, and built directly on the experience gained from the American Memory.

In practice, the staff involved in the American Memory project were, for the most part, those who coordinated the activities in the NDLP. Information Technology Services (ITS) was responsible for computer systems and all aspects of server operations, while the Automation Planning and Liaison Office (APLO) coordinates most activities relating to automation and information technology in Library Services. The Network Development and MARC Standards Office (NDMSO) were assigned to coordinate technical standards related to digital resources.

The materials in NDLP were grouped by collections, and each collection had different characteristics. As of 2005, there are more than one hundred collections in the American Memory Historical Collections. NDLP created a wide array of digital entities for them. NDLP had the same concern for effective access and use of multimedia collections in this complex situation. Providing navigational tools to support effective access to the large body of resources posed challenge. To support this, NDLP developed a range of descriptive elements as content access points which enabled the browsing of subject terms and searching of full text.

The combination of navigational tools supported a variety of approaches to identifying collections and browsing or searching for individual items. Simultaneously, in autumn 1995, the User Interface Team from NDLP was working on a two-year project (1995-1997) with the

Human Computer Interaction Laboratory (HCIL) at the University of Maryland to explore new approaches on user interfaces and organizing the content of the National Digital Library, as the collection grew larger and virtual collections crossed the boundaries of individual digitized collections.

I mentioned in the previous paragraph that the American Memory pilot identified multiple audiences for digital collections. A key discovery was the strong, well-received response in schools, especially at the secondary level. Teachers enthusiastically welcomed digital resources as course supporting materials, so NDLP decided to develop the American Memory Learning Page, which is actually designed for teachers [Lamolinara, 1996]. The Learning Page provided contextual materials, sample lesson plans and activities, and descriptions of the digital collections for pupils from kindergarten up to grade twelve schoolteachers and media specialists in schools.

NDLP employed international initiatives as well as established industry-standard formats for a range of digital reproductions, some of which are XML-based or are becoming XML-based. For example, texts encoded in SGML based on TEI, images stored in Tagged Image File Format (TIFF) files or compressed with the Joint Photographic Experts Group (JPEG). In other cases, the lack of well established standards has led to the use of widely adopted Web-based new streaming technologies, for example WAV and MP3 for sound recordings, MPEG for video compression and delivery, QuickTime for moving images and the MrSID portable image format for map capturing and viewing using GIS software.

It is a challenging task to deliver multiformat objects via the Internet. What NDLP did, led by decisions made by experts in this field, provides an example of "best practice" that the library and information community can follow.

Collections and collaborations

NDLP is a public-private programme that works inter-office within the Library, as well as inter-institution with outside project partners to add to its collections. For example, the Copyright Office had been forwarding regular and helpful guidance in the matter of intellectual property to NDLP, and also had been supporting NDLP by launching a range of information infrastructure rights projects [Levering, 1995]. NDLP and the ITS had been working closely on an on-line access model since the early pilot stage. Even though NDLP has officially come to an end, ITS support of the national digital library exploits multimedia projects and several digital conversion and storage programmes [Library of Congress, 2002]. The National Preservation

Office had been providing conservation treatment for a number of American Memory collections.

Additionally, NDLP works in cooperation with outside institutions such as members of the Digital Library Federation, other libraries and archives, and the Global Gateway to expand its collections. For example, from 1996 to 1999, the Library of Congress sponsored the Library of Congress/Ameritech National Digital Library Competition [Library of Congress, 2003b] along with the first large single contribution from the Ameritech Foundation to enable public, research, and academic libraries, museums, historical societies and archival institutions to digitize American history collections and make them available on the Library's American Memory Website. The Competition also helped fulfil the Library's goal of placing, in cooperation with other institutions, five million items on-line by 2000, the bicentennial of the Library of Congress, making NDLP one of the Library's best birthday "Gifts to the Nation" [Library of Congress, 1999].

In other cases, NDLP has been extending collections through an international collaborative digitization effort, International Horizons, working with the Office of the Librarian, Library Services and Information Technology Services. This effort includes Meeting of Frontiers [Library of Congress, n.d.], the Spain, United States, & the American Frontier: Historias Paralelas project [Library of Congress, 2001b] and many other projects. Other collaborative projects enhanced the national aspect of NDLP such as Civil War Images drawn from the New York Historical Society's archival collections documenting the Civil War and the Nineteenth Century in Print which covered 19th century American imprints in books and periodicals [American Memory, n.d.].

These cooperative projects again indicate firmly that the Library has been addressing through a collaborative programme, combining its own efforts and the creativity of other institutions and communities, the challenges of providing widespread access to the treasures they hold for posterity, in the context of a global architecture for digital resources [Arms, C. R., 1996].

Long-term preservation

NDLP regarded the digital collections as a significant investment, but digital resources cannot serve to replace original materials because there is no way that the original materials could ever be destroyed in virtue of their digitized copies. However, good quality digitized copies can potentially be used for most academic purposes as a substitute for the original. Although the NDLP aimed at access, the Library was nevertheless concerned to manage and preserve the digital archive over time. Some useful preservation methods were identified during the meeting

of the Coalition for Networked Information in 1998 [CNI, 1998]. As the digital versions made for NDLP had the potential to serve as "preservation copies", the Library had been taking many steps during the capture process in the hope of reducing the future cost or need for migration [Arms, C.R., 2000].

Recognizing the importance of long-term preservation, the Library is committed to playing a meaningful role in helping to find creative solutions to the challenge of digital preservation. At the same time, the Library also recognizes that long-term preservation will clearly require a more collaborative approach in the digital realm because there will be more challenging technical, organizational and managerial concerns. Supported by Congress, the Library has begun to develop a cooperative nationwide collection and preservation strategy for digital materials, the National Digital Information and Infrastructure Preservation Program (NDIIPP). Starting in 2001, the Library has hosted a series of convening sessions aimed at gathering together over 140 experts to discuss digital preservation and rights management issues, as the Library must develop the partnerships and cooperative relationships that will be required to continue fulfiling its mission to "sustain and preserve a universal collection of knowledge and creativity for future generations" [Friedlander, 2002].

Historic mission

NDLP afforded the opportunity for the Library along with other agencies and organizations to mount research and development projects directed at real problems of indexing, managing, presenting and storing such diverse collections. NDLP accumulated substantial experience in the operational details of building digital collections that represent both robust systems for today and archives for tomorrow. As the Library's massive multiformat collections grow, the National Digital Library's work still does not end when collections are online. Launching, delivering, preserving and securing digital content and services for Congress and the public in the online environment will be an ongoing challenge. Digital library technology develops rapidly and the Library of Congress will certainly continue through its National Digital Library Program to be the lead organization of a network of libraries and other organizations in developing a national strategy.

Chapter 7

Rationale for Digitization

In this Chapter, a number of tasks relating to effective digital library collections have been identified and are clustered into four major categories below. The creators of the digitization plans and programmes need to be clear about their intentions and objectives with regard to future use. In considering digitization plans, institutions need to ensure that digitized resources can be used for multiple purposes. The views of the case study digital libraries were sought by asking the following questions:

- Why are you digitizing material?
- What types of collections have been selected for digitization?
- What factors influence the content selection and evaluation, from the practical and technical points of view? Does XML technology influence the selection policy?
- What demands (such as meeting preservation needs and increasing access) or institutional strategies (such as reducing costs and attracting funding) will be met by digitizing them?

7.1 Discussion

7.1.1 Reasons for Digitization

Before establishing a digitization project, it would be helpful if there were an explicit statement of the project's vision, policy, or plan that could be used as the basis of a core mission statement. This is to ensure that all parts of the project share the same destination and can measure their accomplishments; this is also useful when dealing with sponsors, owners of data, users and so forth. Several common considerations may influence the plan, for example a guided mission, an institution-wide mandate, the support of the library, university or museum leaders and sufficient funding.

1. Interviews

Perseus

Perseus was born from the vision of the creator, Professor Crane, who desired to assist Humanities scholars in doing research differently. The mission was expanded later to a wider

audience, as indicated by Professor Crane: "We see in the digital environment a possibility to realize more fully the vision which animated the creation of public libraries and this is our long-term goal which is to make accessible, both physically and intellectually, to every human being on this planet the complete record of humanity" [Crane, 1998].

Michigan

The interviewee at Michigan mentioned that the digital library derived from thoughts about the future role of the library in an extensively networked teaching and learning environment. The programme is tied heavily to its parent organization, as the University Library policy states: "The program will focus on developing a networked information environment in support of the academic community..." [Lougee, 1998]. It seemed to me at the time of the interview that this was an early attempt to do what the increasingly common Virtual Learning Environment systems which are being incorporated into most higher education teaching environments in the United States and the United Kingdom are specially designed to do, as I discussed in Section 1.2.5 of Chapter 3.

The Library of Congress

As one of the largest libraries in the world, the Library of Congress has its central mission to serve the Congress. Its valuable and prestigious collections have made it an invaluable laboratory for scholarly researchers. I was told that the development of the National Digital Library initiative was in response to its own sense of mission to the information age. The Library has given itself a mandate to take a leading role in enabling the American people to access information in the digital environment in order to support and contribute to the development of a knowledge-based society. The impression I gained when interviewing was that by taking a leading role in digital libraries in the United States, the National Digital Library initiative could ease the dilemma of extending access from limited readers to the broader public. The original intention of the National Digital Library initiative was to bring together historical documents on Americana to make a reality a collection based on those characteristics which displayed the truly national nature of the United States, and in turn to stimulate users to go back to books to find information to supplement their online research [Billington, 1995b].

2. Analysis and conclusion

Perseus is the exception among the three case studies and many of its features are not applicable to Michigan and LC, since it is the product of a sole creator. This can have its disadvantages as, if that person left the scene or even lost interest, there might be no individual willing or able to continue it.

On the whole, I discovered from the interviews that a strong message lay behind these three Libraries' mission statements, that is the mission of distance education in the context of life-long learning. I believe that this is a key driver in the development of digital libraries because information becomes obsolete more quickly in the information society; it is a fact that learning does not end at the termination of school life but is a life-long experience. On the other hand, we are increasingly facing an actual revolution in the traditional ways of teaching and learning, brought about by multimedia information available online and applied to education. Networked multimedia resources can be a powerful tool in enhancing teaching and learning, as I discussed in Section 1.2.5 of Chapter 3.

Technology changes the method by which people access information. With the help of advanced technology, organized information in digital libraries could be delivered to any corner of the planet and hence fulfil the mission of education. Unfortunately, large areas of the globe do not yet have access to the Internet or sufficient bandwidth to enable access to these rich multimedia digitized resources.

7.1.2 Selection of Materials for Digitization

From my research interviews, I discovered that building and managing a successful digital library initiative could be a long-term and expensive business. It is therefore important to ensure during the selection process that issues of technical feasibility, intellectual property rights and institutional support are considered along with the value of the materials and the interest of the content.

1. Interviews

Perseus

I was told that Perseus' selection criteria have been heavily driven by the project's mission and, indeed, I felt during the interview that I could say by the founder's mission, than by the need to fulfil the requirements of an agency's grant, the needs of an institution, or the expectations of an organization.

I also learnt that Perseus has been increasingly trying to bridge the gap between the traditional academic audience and a wider audience to reflect its long-term mission. This idea thus deflects the selection policy in digitizing materials from the subjects that experts would like to see to those subjects, such as Shakespeare, that would be familiar to the general public. Perseus started out with Greek materials because it was the field most familiar to its creator, and then branched out to all Humanities-related subject collections. It was mentioned that Perseus maintains a particularly strong relationship with partner projects to reach an

integration and synthesis of a variety of formats that share the same topics and are scattered in different locations.

Michigan

I was told that Michigan's selection criteria reflect its strong partnership with surrounding academic departments and institutions, on-campus as well as off-campus. Michigan supports advocacy from the Michigan community for a variety of projects, including museums. This meets the interests of their funders, supports the University Library's current high priority activities such as digital preservation, enables learning materials to be more widely disseminated than they can be in printed formats and also contributes collaborative projects to the development of a critical mass of digital materials in a subject area.

The Library of Congress

As well as evidence from research interviews, I discovered that according to the Librarian of Congress, James Billington [1995a], the acquisition policies of the Library of Congress have placed an emphasis on the materials that are necessary to Congress as well as the immediate concern to the American people; therefore, the digitizing selection criteria focus on manuscript and non-book Americana treasures, such as American literature, history, society, science, arts and music, where these materials are not easily accessible because of their fragility and rarity. Additionally, I was told that the Library holds large quantities of motion picture films and sound recordings, and very probably a large part of them is at a high risk of mutilation because of physical deterioration and the obsolescence of equipment. The Library deems the digital audio-video material to be of long-term value and, based on Library policy, it must ensure the continuation of access to readers. At the same time, some of the digitization undertaken by the Library is under the preservation scheme. Therefore, a number of aspects influence the content selection.

2. Analysis and conclusion

Perseus' selection criteria are different from those at Michigan and LC. In a sense, I think that those of Perseus are more like the expectations of an audience because the organization that hosts Perseus does not have any requirements of Perseus. Perseus has to look to its users for direction. The users are broad based with students of different levels, academics and the general public, and this contributes to what I see in Perseus.

I discovered that Michigan had a wide range of selection policies with strong support from University Library collections, while Perseus was a research testbed and was based in the Department of Classics at Tufts University with very limited University Library support except in the case of the Bolles Collection. Compared to Michigan and LC, the availability of

source materials could be a big issue for Perseus because they do not have access to their own collections. However, the potentially bad situation has stimulated them to establish strong partnerships with outside institutions worldwide; hence, it is the most active in marketing and partnership among the three digital libraries. Among the three digital libraries, LC, on the other hand, is privileged to have massively rich collections which are on a long waiting list to be digitized.

7.1.3 Content Selection and Evaluation

1. Interviews

From all the research interviews, I learnt that both copyright and intellectual property issues present a major hurdle, which was recognized by the three digital libraries as the most influential factor affecting content selection and evaluation.

2. Analysis and conclusion

The issue of copyright and intellectual property protection is not a new one for collection administrators and end users, nor is it an issue that is peculiar to digital collections. The digital era has brought with it a wide range of new types of intellectual creations; nevertheless, digital technologies complicate the negotiation of intellectual property rights, as digital objects are copied and disseminated easily and rapidly at little or no cost to the users. There was evidence from my research interviews that digital collection administrators need to extend significant expertise and time to tackle the copyright and licensing issues before putting the materials online. On the other hand, although digital technology does provide a range of electronic copyright management systems (ECMS), from the user point of view the ECMS may complicate legitimate user access to information. I will discuss access management further in Section 2.2 of Chapter 8.

It is worth noting that the role of the Library of Congress in registering copyright provides a unique opportunity for it to augment its comprehensive physical collections with digital works that might otherwise vanish from historical records. It was largely recognized that the Library required a production-quality system for managing digital objects deposited with it and registered for copyright [National Academy of Sciences, 2000, pp.90-140]. The CORDS was set up at the American Memory pilot stage and laid the foundation for future development on a more robust digital object deposit and register infrastructure, though even in the year 2003, it was still regarded as a pilot and was unable to handle a large quantity of digital materials [Library of Congress Copyright Office, 2003]. One of the most important applications of CORDS was the copyright registration of dissertations through cooperation with ProQuest, a world leading dissertation service based in Michigan. The agreement designated ProQuest's

Digital Dissertations databases as the official off-site repository for more than one hundred thousand dissertations and theses converted to digital form since 1997 and registered electronically. This approach provided a meaningful experimental model for how the Library might handle the digital collection registration.

It was interesting to note that, for the three digital library initiatives, technology had never influenced the projects, but in each case, the project led the technology. When the three digital library initiatives started to develop, they strove to adopt standards and best practices, that is SGML and TEI, as the basis of their technical architectures. They had been consistent with national and international best practices in order to meet user needs and ensure interoperability. The approaches showed that markup technology successfully handled the large quantities of data, and that the technology was principally beneficial to full text searching. Perseus transferred from SGML to XML when XML was emerging. Michigan uses XML in its image collections and has scheduled to transfer their technology from SGML to XML for text creation. LC still uses SGML as a core technology mainly because they had already built digital library infrastructures based on SGML, but, on the other hand, they have been monitoring closely the development of XML-related technologies. Therefore, XML had never been an issue in the selection and evaluation of content.

I noticed that the three digital library initiatives had been actively participating in international activities in advancing the use of "better technologies" (that is, less expensive or compatible with other environments) [Kling, 1999] in libraries and had been eager to collaborate with partners to achieve this end. Perseus was the best case. Perseus had been aggressively seeking expertise, sharing with its European partners for years, examples of such being the cooperation with the Max Planck Institute team at Berlin and the Beazley Archive at the University of Oxford. Meanwhile, I also noticed that the three digital libraries had been monitoring and experimenting with new technologies in implementation, mostly because they took their missions seriously, and wanted to explore every possibility and challenge posed by a digital library. I could conclude that technological innovation can play a major but not the only role for institutions to develop strategies for their digital collections. The true achievement does not necessarily lie in the ability to understand all the expertise in its technical detail, but to put the proper technologies together and integrate them with existing collection management systems.

7.1.4 Demands Met by Digitization

1. Interviews

My last research question related to the demands that would be met by digitization. All three

digital libraries recognized that creating access for a wider audience was the primary digitization demand, since education had been cited as the supreme mission. Preservation, however, is the secondary demand, as in the case of the Library of Congress Digital Audio-Visual Preservation Prototyping Project (DAVPPP). Essentially, the Project has a preservation focus.

2. Analysis and conclusion

Traditionally, libraries have had the responsibility to preserve materials for use by future generations. Today, libraries are facing critical challenges in digital preservation that are widely recognized by library and archive communities. However, to date, digital technologies have not achieved adequate stability for the preservation of analogue materials when the digital version is intended to replace, rather than supplement, the analogue version. I agree that digitization should not be seen as the solution to problems of preservation. I described in Section 2.2 of Chapters 3 and Section 3.3 of Chapter 6 how the Library of Congress, because of its traditional role of leadership, is seriously developing preservation strategies to define national policies and protocols for the long-term preservation of digital materials and for the technological infrastructure that would be required for the Library to achieve this end. I will discuss this further in the following chapters.

7.2 Conclusion

In general, the research interviews indicated that intellectual property rights, considerations of audience and emphasis on the library's unique special-collection holdings affected the selection and evaluation of digital collections. The technology, including XML, was never a consideration but the digital library initiatives led the technology. I saw the three Initiatives had been embracing the challenges brought by creating an integrated collection development policy that could cover possibly multi-format materials, if they felt comfortable integrating the technologies into their infrastructure. Their efforts were intended to create wider access for the possible wider general public to fulfil the core mission of education.

Chapter 8

Realizing the Digital Libraries

This Chapter evaluates the processes which had been undertaken to set up the digital libraries. This includes evaluating the different digitization processes that have been used, the metadata schemas, metadata systems and delivery methods within each digital library infrastructure.

8.1 The Digitization Process

The actual digitization process begins with the preparation of material for digitization, which means considering the applicability of particular techniques against the type of resource and anticipated use, and the device used for digitization in relation to particular types of resources. Quality assurance is a post-process check to see if the decisions made earlier were the right ones. Knowledge about the technicalities behind the conversion process is also involved at this stage. Furthermore, in order to perform seamless search and retrieval in a networked environment, a very important issue is to support a standard way of creating the content and reaching agreement on best practices such as TEI. Although SGML has been implemented and is working stably in Michigan and LC, the use of the new initiative XML is an option to consider. The key questions include:

- What materials are suitable for digitizing?
- Do you use optical character recognition (OCR) to create transcriptions of textual documents?
- What materials are suitable for text encoding?
- Is the text-encoding done by human labour or machine?
- What have you done about quality assurance?
- What are the issues regarding hardware, software and resulting file formats?

8.1.1 Discussion

Before conducting the research interviews, I learnt that up to the year 2002, Michigan had encoded five million pages for its text collection and made more than one hundred thousand

images available online in its image collection. The Perseus database includes more than ten million encoded words and thirty-three thousand images. The Library of Congress originally planned to convert as many as five million of its more than one hundred million items into digital form before the year 2000. By 2001, the Library had digitized over six million items of books and pamphlets, manuscripts, prints and photographs, motion pictures and sound recordings. There was a prediction that, by 2015, digitization of the entire Library of Congress would have been completed [National Library of New Zealand, 2001], but that seems to me unlikely to be achieved, if only because of the mechanical problem of handling so many books rather than because of any digital constraints.

8.1.1.1 Text Conversion and XML/SGML Encoding

1. Interviews

I had learnt from my research that all three examples in my case study used TEI conformant encoded texts in XML or SGML. On most occasions, manuscripts and similar collections will be converted to digital image-only sets, while books and other longer narrative works may require conversion both to an image set and to an XML or SGML-encoded text file. The resulting digital image and text deliverables will be incorporated into computerized presentations that are part of the libraries' programmes.

I learnt from the research interviews that in these projects, in the early days of digitization, text conversion was achieved by data entry. Later, OCR technology was widely used as a means of transcriptions of textual documents, while the limitations in historical materials such as Latin and Greek material, non-standard type fonts, or typography design downgrade its performance.

2. Analysis and conclusion

Although OCR is suitable for many types of documents, it is recognized as being too inaccurate to be used alone, especially for projects in which accuracy is important. Alternatively, rekeying is applied when the highest degree of accuracy is required (the Library of Congress and Michigan specify 99.99% or higher accuracy). In any case, much of the Perseus material is in Greek and Latin. Although some software such as WorldLanguage.com [2004] claims to be able to read and convert Latin and Greek text, and output into XML Output format, it is not clear whether it would be easy with the kinds of materials Perseus deals with, which often have antiquated fonts. In the case of the London project at Perseus, Perseus sought a cost-effective method that could make materials available to the public as quickly as possible. The project relied heavily upon OCR rather than professional data entry because of budgetary and time constraints; yet, Perseus indicated that the implicit labour

(because the OCR-only practice greatly increased the amount of manual post-processing) spent more funds than those explicitly set aside for data entry.

8.1.1.2 Automatic Tagging

1. Interviews

Digital library digitizing projects are often constrained by time and budget, as well as by staff resources, so automatic tagging then is a possible solution for project administrators.

Perseus is the example that has been aggressively pursuing the technologies, taking advantage of them to do things economically. According to Prof Crane when interviewed, Perseus is seeking a long-term collaboration with Johns Hopkins University working on machine translation research. Perseus is pursuing the approach of automatic conversion of a full text database into XML, attempting to determine what quantity of data can be taken out of tags, as far more than normal is done by hand. In the Michigan interview, the respondent reported that they applied automatic tagging technology where they did not need much structure. For example, the MOA project in Michigan was done one hundred percent automatically, but it did not identify anything at the level of phrase or word. LC, I presumed, was not interested in automatic tagging as the digitizing quality is paramount rather than the need for economy with less quality.

2. Analysis and conclusion

Automatic tagging means adding the XML or SGML tags to a document without human intervention. Among the methodologies used are document image analysis and structure and pattern recognition, including identifying structural elements at the page or text-block level, before proceeding on to the recognition of individual characters [Baird et al., 1992], or parsing document "tokens" such as indentation, font size and emboldening to build structures [Conway, 1993] as described in the description of applying SGML tagging based on TEI DTD to text captured from print documents via OCR [Palowitch and Stewart, 1995]. More recent research is trying to explore the possibilities for using intelligent agents for automating these tasks to make them more efficient. For example the collaboration between Perseus and Johns Hopkins University is such a case. However, the technology still cannot reach the expected high level of content quality, so human editing is necessary in the post-processing phase.

My view is that a limited budget can sometimes be an advantage, as for example in the case where Perseus is forced to be creative and address the issue of whether they could convert all the materials automatically and, if not, what proportion could be done automatically. In this sense, I suggest that Perseus has made a substantial contribution to the development of digital

libraries.

8.1.1.3 Digital Imaging

1. Interviews

This topic was not covered much in the research interviews since I was already aware from my research of the practices of the three case studies, also because these are currently the well-accepted technologies for digital imaging.

2. Analysis and conclusion

Robinson's digitization chain theory is based upon the concept that the best quality image will result from digitizing the original object [Robinson, 1993]. This theory suggests that digitization should be carried out with as few steps removed from the original as possible. Every step will therefore mean more links in the chain. If one piece of the link were broken, the entire project would fail. The chain theory could be regarded as the highest guideline when digitizing. However, in some cases a surrogate of the originals will be considered as in the case of large size maps and pictorial source materials, or where preservation needs dictate it, or if a collection consists of a stock of surrogates [Lee, 2001, pp. 64-66]. That was the case with the Library of Congress. LC digitized 35mm microfilm collections produced between 1950-1994 which contained historical materials like manuscripts of American presidents, 19th century magazines and early documents. These historical materials had been microfilmed as a part of the Library's ongoing preservation microfilming effort. James [2003] also suggested that a digital resource should be designed to be as independent as possible from the means of accessing that content. This will help keep the finished resource as flexible as possible, allowing it to change and develop to meet unanticipated requirements.

Frey and Reilly [1999] outlined sets of standards that were considered appropriate for wider access. For example, for resolution, the minimum effective level should be set to 300 dots per inch (dpi), but ideally it should be 600 dpi. Images should be digitized to the highest possible standard, and the resulting files should be saved as uncompressed Tagged Image File Format (TIFF), which is the most widely accepted format for archival image creation and retention as master copy. The digital library initiatives in our three case studies have been following these rules as their best practices. The Ching Digital Image Library project discussed in Section 7 of Chapter 4 followed the same practices. The California Digital Library [CDL, 2001] also outlined a set of digital image standards to meet a more complex digital library environment, but standards from Frey and Reilly still apply to the present digital environment.

The TIFF is a lossless compression, but with advantages for any future use. While TIFFs are

ideal for offline storage, Joint Photographic Experts Groups (JPEGs) are a better solution for online presentation. Unlike TIFFs, JPEGs are the well received file format for Web viewing with the ability to transfer them between systems. This is not only because of their compression capabilities but also their quality. Both TIFFs and JPEGs can transfer to any platform or software system. Also, it is simple to make a JPEG image from a TIFF file as an online viewing copy. The three case studies have adopted this approach as best practices.

8.1.1.4 In-house versus Outsourcing

1. Interviews

I noticed from research interviews that when it comes to deciding whether to do the work in-house or not, the reasons for choosing one option over the other are varied.

2. Analysis and conclusion

One of the advantages of in-house is that libraries can monitor and respond directly to job processing and the outcomes on site, rather than needing to discuss them with external agencies. The other substantial advantage is that it contributes to the expertise of library staff's skills in both hardware and software; this in turn, contributes to the stock of equipment and software. Michigan's initiative is a notable case of an in-house initiative. Michigan used to do in-house digitization and thus has gained significant experience which benefits it in supporting the production service within the library community and, in return, provides constant revenue from the business. Michigan has established a good model of sustainability in the digital library, as we will see when I discuss the sustainability issues of the three case studies in Section 3.1.2 of Chapter 9.

Most collections in LC are digitized by specialist contractors. The main advantage of outsourcing is that the Library does not need to keep a wealth of digitizing equipment, but has people in-house reviewing the digitized materials. Specialized agencies will be able to keep up with the changes in equipment and technique as old ones become redundant. Importantly, they are available with high-level expertise such as microfilm scanning or slide digitization. In other words, they can be efficient and professional. Yet, the disadvantage of outsourcing is the amount of consideration necessary for the detail of aspects of the contract, aspects such as the penalty, should the agency not fulfil its side of the agreement, the risk of damage to materials and so forth. LC stated that they had more difficult detail in contracts which they managed, as there are items with high security at the Library of Congress; thus, the outsourcing firms need to bring in external staff, equipment and expertise in various aspects, including making contracts with government institutions. Michigan now does more outsourcing than in-house because of the advantages I discussed earlier. Perseus always aims

to economize on expenditure, so they consider undertaking every digitization activity firstly in-house, and if that cannot be done for whatever reasons only then do they consider outsourcing.

Digitizing technology has become recognized as the market grows. The three digital library initiatives have experienced no difficulty in finding a suitable outsourcing agency with a competitive price. On the other hand, I suggest that librarians in the digital age ought to keep up with the knowledge of digitizing technology and learn how to deal with the outsourcing agency; thus, they could manage well the complex digitizing tasks. I will discuss the staff management issue further in Section 1.1.1 of Chapter 9.

8.1.1.5 Quality Control and Documentation

The purpose of quality assurance is to ensure as full an information capture approach as possible to digital conversion and to ensure high quality and functionality while minimizing cost [Chapman and Kenney, 1996].

1. Interviews

The three initiatives have developed a set of rules that govern the conversion of a wide variety of collections into digital forms. In the case of Michigan and LC, they considered the resulting document to be an archival form or potentially it could become a long term preservation copy which can be exchanged and used universally without regard for hardware or software type; thus, the quality review has been tightly monitored. Compared with the other two, Perseus is built as a digital library research testbed; materials stored in Perseus are not necessarily to be regarded as materials for future preservation; hence, it is looser in quality verification.

2. Analysis and conclusion

As absolute fidelity to the original source is of fundamental importance, this has been regarded as the priority of the guidelines. The three initiatives have built up a mechanism of on-going quality control and checking. Furthermore, the three initiatives maintain comprehensive documentation for assessment, detailing the data creation process, key decisions and actions at each stage in the project, and have made it available online. Lessons learned from practical experience could then be shared within the community. Another benefit of documentation is that it could help to refine research questions which could be of vital aid to communication in later, larger projects. As the leader in the library community, the Library of Congress is committed to establishing and maintaining standards and practices that will support the development of digital libraries. I note that the Library of Congress as well as

Michigan has been documenting thoroughly its digital library experiences and making these available online, so that they can be regarded as guidelines for use in other projects. Indeed, it appears that libraries, archives and historical societies worldwide have been benefiting substantially from these well-organized documentations.

8.1.1.6 Audio and Video Conversion

A documentary might include moving and still images, speeches or a song. These media have great potential to be incorporated in the digital library services.

1. Interviews and Supplementary Research

I learnt from my research that LC uses MPEG, QuickTime and WAVE files for sound recordings and motion pictures [Library of Congress, 1998d]. Also, in the research interview, the interviewee thought the MPEG family of standardized structure seems to be reasonable practice for dealing with non-text materials; for example, MPEG-7 for audio materials and MPEG-2 for making compressed copy of video. I learnt from the research interview that due to the large amount of non-text materials, the LC DAVPPP was expecting to conduct more experiments with technologies that could provide automatic transcriptions for their millions of deteriorating materials, for example using robotic devices which can operate automatically. But the Project hoped that metadata creation with XML technology would be able to provide efficiency in terms of workflow in order to be able to process as large a quantity as possible in the short term.

Perseus has some DXF files for the Computer-Aided Design (CAD) allowing users to see three-dimensional objects on the Web and some Virtual Reality Modeling Language (VRML) (ISO/IEC 14772-1:1997) files [VRML Consortium Inc., 1997]. This technology allows users to wander anywhere within the models. Michigan holds old films and videos that might need to be preserved digitally. They had been seeking modification of their Image Class system (the name of one of their middleware systems) to support audio and video as one of the new developments in the DLPS Goals 2001-2002. According to my update via email on 17 August 2004, Michigan had not taken any initiative involving the digitization of audio-visual materials, but they were certain that Image Class could manage these materials.

2. Analysis and conclusion

I learnt from the research interview that there were reasonably well-established hardware and software file format options for audio. And there were professional and amateur quality analogue to digital converters for audio. In the case of video, it was recognized that the file format and a file structure for server-based management on compressed video files or

loss-less compressed video files had not yet become well established. The video conversion technology was approximately five or ten years behind in terms of getting systems working out. There are a lot of activities but the outcomes are unclear. This illustrates the difficulties that result from the size and temporal nature of video.

Transforming the great amount of analogue-type audio and video material into digital form brings huge challenges, mainly because of their "size". The delivery over the Internet of these very large files is accomplished by using compression technology. Automatic transcription technology such as voice and music recognition has been widely used in many research projects, though there is not any transcription software giving one hundred percent recognition accuracy. Researchers have recognized that intelligent agents may be particularly useful in the areas of programme research, capture and re-purposing and archiving [Drewery and Riley, 1999]. Agents could be expected to do work such as automatic metadata generation, sort, categorize and index metadata, analysis of metadata and so forth. However, automatic transcription still remains an unsolved scientific problem. In addition, there are differences in the technical requirements to deliver multimedia material to a limited audience as opposed to a large audience through the Internet. Large-scale multimedia material requires complex systems, servers, and wide bandwidth connections.

The XML efforts in dealing with non-text content have progressed rather slowly when compared with text content as discussed in Section 4 of Chapter 2. Apart from the MPEG standard, Synchronized Multimedia Integration Language (SMIL) is an XML-based language for multimedia [SMIL, 2001]. The language is written as an XML application and is currently a W3C Recommendation (W3C Recommendation 7 August 2001). SMIL describes the temporal and spatial layouts of media clips within media presentations. SMIL syntax and semantics can be reused in other XML-based languages, in particular those needed to represent timing and synchronization. For example, SMIL 2.0 components can be used for integrating timing into XHTML, or into Scalable Vector Graphics (SVG) for different purposes, that is, SVG Animation [W3C SVG Working Group, 2005]. W3C Scalable Vector Graphics is a language used to describe two dimensional vector based graphics [Ferraiolo et al., 2003]. It allows for images, text and vector and vector graphic shapes. For example, SVG works with XML DOM and can effect vector animation via scripting.

In reality, although SMIL became a W3C Recommendation early, the SMIL implementation is still not mature, mainly because of the lack of tools. As of 2005, only limited products support SMIL. One such example is the RealOne Platform from RealNetworks which provides support for the SMIL 2.0 language.

Compared with the MPEG family and SMIL, structurally, MPEG-7 is far more focused on audio-visual materials than SMIL, and is not intended to handle the range of materials that SMIL was designed to accommodate. In addition, implementation and industrial support are now better than with SMIL. There are a large number of MPEG-7-related projects being undertaken within commercial enterprises, as well as collaborative government-funded research projects [PERSEO, n.d.], particularly in fields such as broadcasting and digital imaging, which involve the adoption of MPEG-7 and beyond. Furthermore, the implementation of MPEG-21 is growing as time goes on because of the nature of complex digital objects found in digital libraries. This can be seen from one of the LC NDIIPP funded projects [Bekaert and Van de Sompel, 2005].

In general, future research trends could be based on more experiments in using a W3C solution, for example, using XML SMIL language to create interactive content to define a common declarative timing model and integrate this model into XML documents; connecting ongoing work in MPEG with W3C work, for example, combining an MPEG initiative with SMIL along with XML related technologies such as DOM manipulations, XSLT transformations, hyperlinking and so forth.

As for the proprietary data format, Internet streaming media are poised to become global media with many well-accepted Web-based technology options available, but it can be a challenging process to try to select the appropriate format [Beggs and Thede, 2001]. The Library of Congress expects to create and make available resources using the streaming media version to the general public as network bandwidth to homes and schools increases and the technology on typical desktops improves [Library of Congress, 1998d].

8.1.2 Conclusion

Heterogeneity in the digital library presents a challenge that will need to be met. The many advantages of the emerging digitizing technologies are becoming clear. New technology is in constant flux and will continue to provide better solutions for digital reproduction. It is fair to say that text-tagging technology is mature but would need more powerful authoring tools in the market. Automatic tagging technology is still a research area that needs a great deal of effort to be put into it. When SGML was the markup language most used, probably there were research projects to convert OCR into tagged SGML text. It is unlikely that this route would be followed now, since XML technology is more advanced and any work on automatic tagging would be undertaken in the context of XML. Artificial intelligence could probably partially replace human intelligence, as information technology advances so quickly, more quickly than humans could imagine; nevertheless, it would be difficult for human intelligence

to be perfectly imitated by the machine, particularly in anything related to the area of intellectual justification. I think that the main point to concentrate on then would be raising the percentage of retrieval. XML could possibly help with the structure, as it provides great advantages in defining a strict structure and in granularity. Perseus shows that it is a technology-oriented research testbed. The Perseus team regards XML technology as a challenge, and explores the extent that XML can help them to build a digital library.

The library community and research institutions would thus need to work cooperatively to contribute to the efforts of both parties to establish standards and procedures for the best development of digitization. Digital librarians would need to have the knowledge to handle the task of mass digitization, digital storage and digital preservation, providing access to and retrieval of digital information.

The major problem in managing audio-visual materials is the lack of standards as discussed in Section 4 and 5 of Chapter 2. In this sense, DAVPPP at the Library of Congress could provide the guidelines as best practices in the library and information community.

8.2 Infrastructure

8.2.1 Metadata and Metadata Systems: XML Application

The concept of sharing resources in the digital environment is essential. In order to perform seamless search and retrieval in a networked environment, a very important issue is to support standard ways to create and structure their content by reaching agreement on metadata standards. At the same time, it is necessary to use more intelligent search tools which will deliver better quality results. Therefore, the use of new initiatives such as XML is an option that should be seriously considered, since XML can assist with structuring content. Perseus is the only initiative of the three using XML as a core, and Perseus demonstrates the potential impact of XML in digital library development. Considering all these aspects, questions could include:

- What metadata formats have been implemented for digitization?
- How do you manage the consistency of metadata formats?
- Does your metadata system use XML technology?
- What is the level of detail in the metadata tagging?
- Do existing metadata schemas work well? If not, what do you do about this? Do you

create your own?

8.2.1.1 Discussion

In the following sections, I will discuss how the three leading metadata initiatives, MARC, TEI Guidelines (Header) and Dublin Core, were deployed in my three case studies. I will also discuss from the evidence gained from the interviews the advantages and disadvantages of these metadata schemas when applying them in digital libraries, and indicate related XML-based new efforts or approaches that could potentially be used in the digital library environment. Finally, I will attempt to indicate solutions for the complex metadata challenges that are being encountered in the development of digital libraries which involve adopting XML technology and the newly released metadata system METS.

8.2.1.1.1 MARC

Large-scale digital libraries available through the World Wide Web now make available vast quantities of material independent of the physical location of the originals. A large portion of these materials are traditional library materials, for which MARC was invented, but records are available for digital materials, for instance those included in the CORC project shared-cataloguing database [CORC, 1999].

1. Interviews

Perseus

As I discussed in Section 3.2 of Chapter 6, Perseus has no support from the University Library, and does not have input from university librarians, so they do not use the MARC record.

Michigan

Michigan has been working with the MARC records held in the University Library, but has created more description metadata, since MARC itself cannot provide enough information. Then Michigan converts the information to SGML-based TEI Header and adds extra information where appropriate for electronic applications. Michigan is expecting to move to a Unicode compliant library management system, Ex Libris, which has XML capability. Michigan staff hoped that XML technology will soon be adopted in every possible area including cataloguing, and it is particularly important that a solution should be found to the Unicode issue which has troubled them since they started to digitize.

The Library of Congress

Most of what the Library of Congress holds is described in MARC records. They use MARC records, as well as records which they just call generally non-MARC records that are usually prepared in an Access database or some other simple relational database. The format that they use has been used since before XML existed, but it is XML-like, using tags and angle brackets. What they indexed for American Memory is a combination of MARC records and this kind of tagged record. What users are searching for are both kinds of records. Also, in some cases, the metadata is embedded in a full-text file to enable full-text searching.

2. Analysis and conclusion

Problems with MARC

My view is that MARC is not without its problems as analyzed in this section. LC NDMSO has had to work hard, ensuring that MARC remains compliant with new technology. The effectiveness of searching can be significantly enhanced through the existence of rich, consistent metadata as found in MARC. However, MARC was not designed to support direct bibliographic access, since it is a machine-readable format not intended to be read immediately by humans. That is why the MARC format is regarded as being complex and too specialized to be used outside library applications. Now, digital libraries are seeking to implement something similar (hence TEI Header-MARC-Dublin Core mappings) because the MARC record itself cannot provide enough information due to its inflexibility. LC and Michigan did exactly this. Today, a vast amount of the information available on the World Wide Web is not packaged in a format that is well suited to description in the MARC format. Although there had been a certain amount of effort under the auspices of UNESCO with the linking techniques developed in the Common Communication Format (CCF) in adding extra-rich linking ability on MARC in SGML [Guittet, 1985], the MARC format is not well suited for the digital library in general to capture multidimensional relationships among bibliographic records, such as hierarchical relationships and multiple versions related to a single bibliographic record.

Librarians in general have become more conscious of the lack of hierarchies because users have found catalogues to be deficient in making links between different manifestations of a work, hence the interest in Functional Requirements for Bibliographic Records (FRBR) which has been developed by IFLA [IFLA Study Group on the FRBR, 1998]. Some ILMS vendors have been keen to 'FRBRize' (pronounced 'ferberise') their products, for example VTLS [Chachra, 2002] which, as Chair of the IFLA Working Group on FRBR Le Boeuf [2003] reported, would not be possible without encapsulating MARC in XML.

XML efforts

Firstly, the MARC XML DTD started to be used in year 2001. From the research interview, I learnt that actually NDMSO took the old SGML DTD and turned it into XML DTD, using an already available converting tool and did it on the fly. XML MARC provides a more flexible structure for data conversion, not reliant, as is MARC, on fixed field data. Using XML/XSL technology, a librarian can create native bibliographic record files and publish them in different formats for various purposes, such as a public OPAC display format, with multi-scripts display; bibliographic records can be viewed directly by the Web browsers and library systems without further conversion.

The flexibility of XML and its support of Unicode make it a suitable basis for future standards for bibliographic and authority data and could solve problems found in MARC cataloguing in the handling and linking of authority metadata. Establishing a central repository of authority metadata in XML format equipped with an XML linking ability would mean that it is not necessary for each library to build its authority control database locally. Libraries could retrieve via URL (each authority record is an XML file) the appropriate metadata, and thus perform more efficient and economic work.

Secondly, from the research interview, I also learnt that LC NDMSO has developed a new MARC XML schema, called "MARC XML Slim", which is a very different approach from element-based MARC. It is not as large as MARC XML DTD but small, because it uses attributes instead of elements. MARC XML Slim has very few top-level elements, but uses attributes to tell the tags what the indicators are.

Thirdly, at the same time the LC NDMSO is experimenting with the human-friendly XML schema language, based on mini-MARC, the Metadata Object Description Schema (MODS) [MODS, 2002]. From the research interview, I learnt that MODS regroups some elements from MARC in a more logical way, but it is still comparatively tied to the semantics of MARC. MODS uses language-based tags rather than numeric ones, so users would be able to create MARC records in a fairly straightforward way. MODS meets the need for institutions which are looking for something that is less complex than MARC and could be used in XML-based systems rather than in a system only designed for the library to use. MODS is now available for experimentation and evaluation. Because of the ease of MODS, MODS along with the new initiative METS could potentially be able to provide an infrastructure for metadata in the library that complex MARC could not reach. A Library of Congress Web preservation project, MINERVA (Mapping the INternet Electronic Resources Virtual Archive), is a good testbed for the combined use of these two XML initiatives [Library of Congress, 2004].

MARC is the metadata schema favoured by librarians because it is the main standard that they use in their traditional libraries and they are familiar with it. Those digital libraries which are situated in a traditional library will always be based on MARC so long as MARC continues to be such a dominant library standard. With the constant XML efforts from NDMSO, keeping MARC competitive against other metadata formats, I estimate that MARC will still be viable in the digital age.

8.2.1.1.2 TEI Guidelines (Header)

The TEI Guidelines provide a mechanism for identifying metadata as well as for encoding the content of an electronic text. The use of TEI for content is also discussed in this section after TEI Header.

A. TEI Header

1. Interviews

Michigan and the Library of Congress

For descriptive metadata of the volume, I learnt that LC and Michigan are populating the TEI Header with data from MARC records, but this does not apply to Perseus. Because these are high quality records, it enhances the quality of the TEI Header and enforces the consistency of the records. It is often the case that digital library text conversion projects following the TEI Guidelines use a mapping between elements in the Header section of the marked up file to fields in a MARC catalogue record. In some cases, the TEI Header is derived from the catalogue record; in other cases, a catalogue record is derived from the Header. Structurally, the formats are very consistent. In the case of Michigan, Michigan applies an ID system (called 'Notis ID'), which is adopted from the monograph catalogue number. Michigan uses this in the digitization process without losing the records. With this ID system, Michigan can allocate the metadata throughout the process, no matter whether the digitization is undertaken in-house or by outsourcing vendors.

2. Analysis and conclusion

Managing a growing number of data formats and metadata schemas in an integrating digital library service environment is a challenge. Utilizing good consistency of metadata formats like TEI Header facilitates resource discovery in the digital library. In practice, the digital library tends to operate on a case by case basis by tailoring its various mediating services for individual networked resources. The consistency will be achieved through rigid adherence to prescriptive data and metadata standards. Meanwhile, the digital library needs to consider how to implement such standards locally to support the searching, retrieval and other mediating online services.

In general, TEI scheme plays a key role as a quality control mechanism during resource creation management and as a source of rich information for use in resource discovery. The three libraries shown are useful examples of good practice metadata techniques that can be applied within an SGML or XML framework. In the next paragraph, we will discuss, as one of the research questions in this section, the role of TEI in quality control in my three case studies.

B. Consistency of Metadata Formats

1. Interviews

Perseus

According to its literature, Perseus supports the knowledge-based digital library, and has been developing a generalizable toolset to extract structural and descriptive metadata from documents and deliver document fragments on demand, to analyze linguistic and conceptual features and manage document layout [Smith et al., 2000]. Perseus has added a couple of extensions on Vanilla TEI DTD (that is, TEI DTD without modification), and as long as Perseus has valid DTDs, they can handle document management at any level that fits the Perseus architecture. Although XML Namespaces encourage markup reuse and minimize duplication of semantic structures, it is unlikely that all marked up documents will eventually use the same structure of DTD. The Perseus system, therefore, creates structural mapping, which does not require modification in either the documents or the DTDs. This allows the system to create partial mappings between the elements in a DTD and to abstract structural elements. The system next produces an index (termed a lookup table, LUT, in Perseus) of the elements, which is then mapped. Identifier attributes on XML elements, such as ID or N, are also indexed. The occurrences of each structure within the document are sequentially numbered to enable resource discovery and visualization. The benefit of this structural mapping effort is that it happens with external metadata and it is not necessary to modify either the document or the DTD [Smith et al., 2000].

Michigan

Michigan is satisfied with the TEI Lite (a subset of TEI) in marking up the structure of text. In the case of American Verse project, the TEI Lite DTD has proven to be more than adequate for work to be done completely; only minor deficiencies (that is, where figures may be located) have been encountered with the TEI Guidelines. This gives evidence of the benefit of the Guidelines that they allow for description of the content of a document to be accomplished at any degree of granularity. In some cases, TEI Lite does not represent the reality very well, especially for the earlier books. In this case, Michigan has its own modified TEI-based DTD. Like Perseus, Michigan rarely creates their own DTD because they found it easier to derive from TEI-based DTD. This gives evidence to another benefit of TEI that it is

an extensible, flexible and robust scheme for use in the digital library and its primary purpose is to add value to an SGML or XML-encoded document. On one occasion, Michigan tried to create its own DTD, and thought they made a lightweight bibliography DTD for the English bibliography, but the outcome was not what they expected. In addition, Michigan uses the "vendor DTD" which is TEI-based DTD streamlined down for the outsourcing vendors. The vendor will then follow the subset and do the markup. Michigan converts them to TEI Lite when they return.

The Library of Congress

The American Memory Document Type Definition (AMMEM.DTD) is developed to accommodate a broad range of materials by conceptualizing a generalized Humanities text, rather than seeking to describe specific document types and subtypes, or text genres. LC implemented a TEI-based DTD in 1993 [Friedland, 1998], but because the TEI was incomplete at the time that AMMEM.DTD was developed, the Library created its own simple models for some elements [Library of Congress, 1998a]. Over a period of time, the American Memory DTD took into account feedback from its actual use, development and refinement. From the research interviews, I learnt that LC also anticipated making future revisions to the American Memory DTD to accommodate the emerging XML specification.

2. Analysis and conclusion

I think one of the advantages of using XML could be that it is a convenient way to validate records using a DTD or schema as XML technology grows. LC has developed various well-organized consistency checking guidelines that can be applied to text, image and all digital formats. In some cases, the consistency may be imposed because the records actually used are constructed from a database, and this imposes some consistency on entering the data.

C. The TEI Guidelines

The TEI Guidelines provide not a single DTD but an environment for creating many customized DTDs. TEI Lite is one such customization. From my research, I learnt that during the American Memory document analysis, LC library staff and the SGML consultants were surprised by the close fit between the requirements of American Memory and the Guidelines [Lapeyre and Usdin, 1996]. This explains why a major electronic text creation project in the academic and research context would at least start by first considering the use of the TEI scheme.

A set of recommendations with five encoding levels of electronic text was generated from a meeting sponsored by the Digital Library Federation for libraries using the TEILite DTD v1.6 [DLF, 1999]. In other words, it is only when librarians are using TEI that these levels are

recognized. There are many different library text digitization projects for different purposes. Therefore, the level of detail in the tagging varies. The meeting recommends that texts at Level 1 are created and encoded by fully automated means, using uncorrected OCR of page images, or exporting from existing electronic text files. Level 2 texts should carry minimal encodings which allow for keyword searching, linking to page images and identifying a simple structural hierarchy to provide greater navigational possibilities than Level 1 encoding. Level 3 texts can stand alone as text without page images and, therefore, can be uploaded, downloaded and delivered quickly, and require less storage space than digital collections with page images. Level 4 supports greater description of function and content. It allows for flexibility of display and delivery, and sophisticated searching within specified textual and structural elements. It combines the broadest range of uses and audiences.

1. Interviews

Perseus

Perseus is focused on level 4 at the lowest, because their inter-textual linking technique relies on correct and rich tagging in the source text. Nevertheless, Perseus is looking for level 5, which requires subject knowledge, and encoding semantic, linguistic or other elements beyond a basic structural level.

Michigan

Michigan varies tagging practice, since they have been working on digitization projects for different purposes, but in general most text encodings conform to level 4. For example, in the case of strict preservation digitization, the primary motive is to image the pages to create preservation copies. Because there is a MARC record for the original from which the digitized copy is made, the descriptive metadata from the MARC record will be very rich, but details of the SGML tagging will be light, such as a tag for each page. In the American Verse project, the tagging was done in-house in order to create the collections of their holdings, and the tagging was very rich. It identified all the structures that can be identified from the formal publications, at least to the extent that they were known to graduate level students in English or History (they were doing the markup). For the Early English books online, materials are encoded to a level that allows further work by the text creation partnership members.

Michigan has normalized SGML tags, thus XML-like SGML structures. Because Michigan wants to have things very regularized, they all have attributes in DTD order, and all elements were normalized to upper case, which is a slight problem for XML. Since XML is case-sensitive, certain special words must appear in a particular case. In general, the keywords that relate to DTDs, for example DOCTYPE, ENTITY, CDATA and ELEMENT must be all uppercase. On the other hand, the various strings used in the XML declaration, for example

xml, version, standalone, and encoding must appear in all lowercase. However, the programs or macros are not difficult to write to do the conversion. Therefore, Michigan expects that it will not be expensive to convert SGML to XML. As learnt from my follow-up email correspondence in 17 August 2004, Michigan's search engine has now become Unicode compliant; however, there is a problem with the character entities as Michigan has a large number of special characters in English that are not necessarily in the Unicode set. Michigan hoped that the TEI-Non-Standard Characters Group could meet to find ways forward for these problems.

The Library of Congress

LC focuses on level 3 in American Memory. At the time of interview, the LC American Memory DTD (conforming to TEILite DTD) is still SGML. There is an expectation that American Memory DTD will create an XML version; however, there has been no pressure to do that because all the existing tools work with SGML, and they know that transferring will be fairly straightforward because there is not much that is likely to conflict with XML.

2. Analysis and conclusion

Problems with the TEI Guidelines (Header)

Firstly, in the digital library, some documents are available online only as page-images, others only as searchable text marked up in XML or SGML, and others in both forms, image and text. Links from the marked-up texts to page-images and illustrations pose a problem. Projects deal with this in various ways.

Perseus produces the FIGURE element with image ID number attribute for the text caption accordingly, and then the image program will do the actual work, generating HTML using IMG element and SRC attribute. LC uses SGML entity references to provide a level of indirection. Links are marked by one of three elements (corresponding to pages, illustrations, and tables) and identified in an entity attribute that names the external entity within which the graphic image is stored. Michigan has not found a feasible solution to reconnect the texts and the images, and has found it harder than it should be in practice. So in fact the books are online, but the images are not.

The Ebind (Electronic binding) DTD is created to provide structural metadata regarding digital library objects within an SGML encoding format where TEI could not properly solve the problem [Ebind, n.d.]. Ebind is based on TEILite with the ability to include transcription of page images and metadata, but does not allow for true TEI encoding of the transcriptions. One fundamental difference in concept between Ebind and TEI is that Ebind focuses on the physical structure of a document while TEI focuses on the intellectual structure. In TEI, there

is no element which can contain a page. TEI describes the hierarchy of the page implicitly through the use of the <pb> (page break) tag which is an empty tag only showing the start of each folio. In Ebind, all pages are encoded within a <page> element with a "seqno" (sequence number) attribute. This allows one to gather together a variety of information associated with individual pages, such as textual transcription, "raw" OCR or keyed. Ebind also claims that they are simpler than TEI to use. Many of the requirements imposed by TEI are "loosened up" in Ebind, and so it is suitable for use in a high-volume production environment.

Instead of attributes, the Roman de la Rose project made a custom extension to include a new tag <image> within the TEI <pb> tag, no longer strictly an empty tag but still functioning largely in the same way. Actually, this is the concept of how TEI was initially designed, that is users can define their own custom tags, rename TEI elements to a form that is more acceptable to the local applications, define a new base structure for their information, undefine existing elements, modify content models and so forth. The new <image> tag includes attributes for the first and last lines of the manuscript page. The <image> tag also includes a sequence number attribute, borrowed from the Ebind DTD. Thus the <image> tag contains links to the page images, numbered sequentially in the new seqno attribute of <image> [Roman de la Rose Project, n.d.].

Secondly, the XML or SGML TEI DTDs in the three case study initiatives were originally developed for validating XML or SGML-encoded metadata against the Libraries' specifications. Validation is the main benefit of using DTDs, and this supports the consistency of metadata formats. The DTDs have actually been around for quite a long period of time and are likely to be around for a while, since there are plenty of legacy documents that still rely on them for their structural definition; the documents in the three Libraries are such cases. However, as I discussed in Section 2.3.3 of Chapter 2, the XML Schema offers much more functionality and versatility than DTDs. Although the latest version of TEI (P5) is based upon the XML Schema language, it is a very newly-released version that certainly needs substantial feedback from institutions who actually implement it in their digital projects.

Thirdly, the TEI Guidelines precisely design the <fileDesc> component of the TEI Header in supporting the bibliographical description for the library community. The reason for this is that there is a feeling in the library world that a record for a bibliographic item should be created only once – by the first library that comes across a resource – and thereafter reused by other libraries. However, in the case of attempting to use TEI bibliographic data in a MARC record, there are problems. In the real world, the automatic conversion of the components of the TEI <fileDesc> on to corresponding MARC fields is difficult; a certain level of human labour and human intelligence is required [Burnard and Light, 1996]. The Oxford Text

Archive sponsored a meeting in autumn 1997 and one of the outcomes of this was a recognition that only if data in the TEI Header were controlled to the same extent as traditional cataloguing would the data be of value to libraries. It is clear to me that libraries require electronic texts to provide as good a quality cataloguing as they provided in their own catalogues in order to provide consistency across the catalogue. This begs the question as to who provides these data and the economics of providing them. These are ongoing issues, although there has been considerable progress if only in an increased recognition of the problems particularly within the library community [Caplan, 2000].

Fourthly, TEI Guidelines are not able to manage non-standard characters as Michigan found. This problem will be a cost to Michigan when switching to XML as Unicode does not cover this either.

XML efforts

Firstly, the XML version of the TEI Guidelines provides a compatible version of SGML suitable for use on the Web [Sperberg-McQueen and Burnard, 2002]; XML is simpler to parse than SGML, enabling lightweight, inexpensive software of all kinds. Perseus started out with SGML before XML existed, and then migrated to XML. Technically, Perseus still works on SGML encoding because it is easier to type, and does not have to put in tags and quote attributes and so on, using existing tools, and then convert SGML files to XML.

The adoption of Unicode as the underlying character set for all XML documents made it possible for computer systems to interchange documents with generally agreed universal character sets. The XML version of TEI Guidelines supports Unicode, however, there is always the possibility that characters needed are not defined in Unicode such as in the Michigan case. Thus, I suggest that entering and displaying non-Unicode characters is an issue that needs to be addressed. TEI is working on some approaches to deal with special character sets. For example, scholars working with ancient Chinese texts might find TEI/XML KanjiBase useful. It provides a common reference to all characters missing from coding systems currently in common use, allowing for viewing and printing of the character [Wittern, 1995; Wittern, 2000]. Modern Chinese characters are covered by Unicode, and therefore are available in XML.

Secondly, Hockey, who is a TEI senior member, suggested that a possible solution for links from the marked-up texts to page-images was to use attribute IMG within <pb> tag [Hockey, 2003]. In that case, users can put the file name of any object format as a value of attribute IMG. Furthermore, with TEI Guidelines, it is very possible that institutions define an XML version TEI-based markup system to deal with non-text materials by taking advantage of the

XML smart linking mechanism. The institutions currently using the TEI P3 can benefit from the new TEI P4 edition which is expressed in XML. This can be processed and maintained using readily available XML tools instead of the special-purpose software originally used for TEI P3 [Burnard, 2002].

Thirdly, as I discussed in Section 2.3.3 of Chapter 2, XML Schema is an effort of W3C to aid DTD in the XML world. XML Schema aims to be more expressive than DTD and more usable by a wider variety of Internet-based applications. The XML Schema is gaining ground gradually. MODS, the Holdings Schema of Z39.50 (version 1.4, November 2002) [Library of Congress Z39.50 Maintenance Agency, 2002] and the new initiative METS are the best examples of such a tendency; and these are all expressed using the XML Schema.

Fourthly, the non-standard character problems may be solved soon: work has been going on in ISO to ensure that all characters required by early printing and manuscripts will be included in Unicode [ISO/TC46, 2004]. XML Systems with Unicode compliance are expected to provide the solutions for digital libraries with obsolete characters.

The TEI Guidelines are successful not only in the TEI Header as a metadata schema, but also in the mechanism for transcribing and encoding documents. The Guidelines are a large and complex specification and they support different DTDs suitable for different research purposes. We have seen the development of the TEI organization for years; they participate actively in XML development with the motive being to keep TEI Guidelines competitive for adoption in a networked digital library environment. However, Price-Wilkin [2002] from the University of Michigan questioned: Does TEI serve the resulting range of needs adequately? Should TEI create a broader framework for digital libraries? People argued how TEI could remain a vital part relating to other kinds of digital information, such as sound files, maps, compound documents and mixed format repositories which have become the everyday business of digital libraries. Indeed, dealing with the multiple formats of the context of digitizing materials, it seems to me to be reasonable that digital library initiatives need to seek economic solutions, not only with regard to finance, but also in efficiency which would allow them to make progress in every possible area considering the state-of-the-art technology.

8.2.1.1.3 Dublin Core

Although Dublin Core is not the core metadata system in the three case studies, they are making use of some of the Dublin Core efforts such as applications in OAI serving as an interchange format for harvested metadata from databases within and outside library systems, and applications in Z39.50 serving as a basic profile for the information retrieval protocol for

cross-domain discovery. In the following paragraphs, I examine how these technologies are implemented in the three case studies, and the possible XML efforts.

1. Interviews

Perseus

Perseus uses Dublin Core for some records but the main technology is TEI. Perseus provides an OAI harvesting and searching service. When performing searches for documents that are Web-accessible, Perseus creates two links: one directly to the outside page, and one with added Perseus links. As with texts in Perseus, these links point back to related resources in the digital library. Users can also exclude OAI results by de-selecting the "include external sites" option. Perseus does not provide a digital library operation service like Michigan or LC, and so it does not have Z39.50.

Michigan

Michigan uses Dublin Core in XML (it validates as SGML but the big tree is XML) as the set of common fields, mapping a wide range of metadata formats held in individual institutions for cross-collection searching (they are mainly image collections). In other words, the user can either search an individual collection by its original fields or search fields grouped by Dublin Core. In this case, individual institutions do not need to change their working practices and they may have their collections online, as different museums and different department have different metadata needs and different working styles. I see the advantage of simplicity in Dublin Core here serving as a simple alternative for library resources discovery.

Michigan provides digital repository resources through OAIster which constitutes an interlinked network including more than 230 institutions. OAIster makes use of Michigan XPAT Unicode functionality, when searching in OAIster, users are searching a wide variety of collections from a wide variety of institutions. These institutions have made the records of their digital resources available to Michigan, and Michigan has gathered and aggregated them into the OAIster service.

From my research, I discovered that Michigan provides access through their library system, MIRLYN, to other catalogues in other academic libraries via Z39.50.

The Library of Congress

LC has harvested cataloguing data (electronically) to create MARC records and share their experiences with active and potential users. At the time of interview, LC was making their data available through OAI in 4 formats. These records start as MARC records in the MARC communication format. Then they are converted to a similar schema called OAI-MARC,

which existed before the LC NDMSO ever developed any XML-based structure. The third format is OAI-DC, which is mandatory. This uses a mapping developed by the LC NDMSO. The fourth format is MODS. It is intended in some way to be MARC-like, though, because it uses words for tags rather than numbers, it is much friendlier to the human reader. All the interactions in the protocol and the responses of the protocol are based on XML.

The Library of Congress is the maintenance agency for Z39.50 and the technology is implemented in their library service infrastructure.

2. Analysis and conclusion

Problems with Dublin Core

Firstly, although Dublin Core is renowned for its simplicity and flexibility, these could be disadvantages for the digital library. Dublin Core was established as an international standard, but it seems to me that it is so flexible as to be hardly a standard. The fact that it is not strongly defined makes it supposedly easy to use: easy to create the metadata, but the lack of specificity means that the item described might not be easily found, since too many records are recalled. Western names are a case in point; they are usually reversed in metadata. Nowhere does it say in Dublin Core that the name should be reversed from normal order. Additionally, there is a certain amount of ambiguity, for example, creator and publisher are not always unambiguous. When metadata produced according to Dublin Core is extended into systems alongside data created in MARC, there can easily be a lack of compatibility which would be removed if Dublin Core were more precisely specified as cataloguing rules are. However, Dublin Core is kept simple to encourage creators of digital resources (who are not librarians) to use it. But its simplicity could be its downfall.

Secondly, one possible disadvantage of OAI is that the technology has not yet matured and the carrier format for Dublin Core may be XML, SGML, HTML or indeed other formats and whether these can all be harvested effectively is questionable. Certainly, if XML was adopted as the universal format for documents in OAI, OAI would then be more effective because it is based on an open environment and the situation would be easier for the OAI service provider or repository. Another possible disadvantage is lack of consistency in Dublin Core elements caused by the lack of specificity in the standard. A third disadvantage is that the OAI is an HTTP-based protocol but not a Simple Object Access Protocol (SOAP) based Web Services. I will discuss Web Services, an XML effort, in Section 2.3 of this Chapter.

Thirdly, Z39.50 has an active implementers group including experts from North America and Europe. In the United Kingdom, there was a consensus of the need to network library OPACs, in particular in academic libraries, leading to national Large Scale Resource Discovery

systems using Z39.50 to create distributed union catalogues [MODELS, 1999]. Although Z39.50 supports Dublin Core marked up in XML as a transfer format for records [LeVan, 1998], the infrastructure is still not properly defined.

XML efforts

Firstly, Dublin Core has great advantages of simplicity and flexibility; yet, these could be disadvantages as far as meeting the complex requirements in digital libraries is concerned. The XML Namespaces might be one of the possible solutions to the disadvantages of Dublin Core. Therefore, as we have seen in Section 2.1 of Chapter 5, it seems certain that the evolution of XML/RDF will continue to stimulate the development of an underlying data model for Dublin Core.

Secondly, an initiative of Library of Congress and others on Z39.50 has been developing; it is called ZING: "Z39.50 International Next Generation" (earlier name was ZNG: "Z39.50 Next Generation and before that it was called for a short time ZML: "Z39.50 over XML"). ZING aims to promote investments in existing Z39.50 service and specifications and facilitate interoperability with the evolving new Web technologies such as XML and SOAP [Jørgensen, 2000; Denenberg, 2002]. Therefore, it is most likely that the next generation of Z39.50 will be based on XML technologies.

In practice, Z39.50 can be seen supporting many of the XML family of technologies making it work in a flexible and extensible XML environment. For example, in a Z39.50 environment, application programs may access XML structures via the Document Object Model application program interface; the XSL transformations can handle the displays; users can search XML information via XQL search language; the document attributes and relation types may be defined by XML Namespaces (for example Dublin Core format) in RDF structure; the infrastructure can use XML-based SOAP as a message exchange model [Jørgensen, 2000].

The Dublin Core is a much discussed set of metadata elements within the library community and beyond and is increasingly regarded as a metadata schema that can provide a useful basis for general purpose networked resource discovery. The Dublin Core Metadata Initiative has been evolving and building consensus through a series of workshops and a wide range of interest of working group activities [Weibel and Koch, 2000]. Active discussions have highlighted the interests and expectations of different communities. As adding functionality to meet the needs of domain-specific applications is one of the Dublin Core's main goals, the development of Dublin Core Qualifiers plays a key role in supporting the broad interoperability.

Z39.50 and OAI share the same motive behind the Dublin Core interoperability efforts, that is to enhance broad access to valuable resources via the Internet. It is my view that they could be built under the XML environment and do things efficiently as the technologies and tools grow, thus achieving the aim of sharing the scholarly resources. The two approaches add value to the service of digital libraries.

8.2.1.1.4 METS: an XML Effort

I did not ask a question about METS in my questionnaire and it arose only in the research interview with the Library of Congress as at that time METS was beginning to be discussed in the library and information community. Afterwards, I did further research and corresponded in August and September 2004 on the use of METS with the three digital libraries.

1. Interviews and Supplementary Research

Perseus

Perseus will be creating full METS records when in the process of transferring their digital objects to the Tufts Digital Library, which uses the METS descriptors.

Michigan

Michigan was expecting to use METS as an interchange format for different digital objects. According to DLXS News dated 20 March 2003 on their Website [DLXS, 2003a], at the first meeting of the Michigan DLXS User Group, held in November 2002, members gave the highest number of votes for generic support of METS in DLXS. Also, according to my enquiries in 17 August 2004, Michigan was in the process of hiring a new programmer to work on METS integration.

The Library of Congress

METS will be implemented in LC's central digital repositories. In addition, the Library of Congress DAVPPP is investigating the MPEG family of standardized structures for the compression and storage of bit streams and video. One possible way of doing this would be wrapping METS around in MPEG-7 along with other metadata format structures, for instance the metadata standard from the Audio Engineering Society (AES), which is an emerging metadata structure for technical information about sound files. Furthermore, The Library of Congress is adopting METS and Open Archival Information System in the preservation project as METS is a package for metadata and this has the similar role to that of information packages as defined in the Reference Model in the OAIS [Rick, n.d.]. METS will continue to be used in more new projects. An example of such is the 'I Hear America Singing' (IHAS) project, which uses METS for digital objects.

2. Analysis and conclusion

Problems in metadata

Firstly, managing the complexity of digital collections, or even single digital objects such as online books, may seem challenging. Without administrative metadata, it is difficult to discover who can scan the item, what copyright restrictions exist, to which collection the digital object belongs and so forth. Without structural metadata, it is difficult to understand how a digital item is organized, for instance keeping image and text all together in the digital work. Without technical metadata relating to the digitization process, it is difficult to understand how accurate a reflection of the original the digital surrogate is, for instance researchers may want to know the technical characteristics of the digital object and what has happened to it, since it entered the collection.

Secondly, most of the metadata schemas were developed within specific communities, though it could be argued that Dublin Core began with the library community but was quickly adopted by the world at large. Of the three formats I discussed in this section, TEI Header, MARC and Dublin Core, share common data elements, but they are separate technologies with different reasons for their existence because they are serving particular communities. For example, Dublin Core is for those who find TEI Header and MARC too time-consuming to use. The pieces of information they deal with exist at different levels of granularity. The three cannot be perfectly aligned with each other on an element-by-element basis.

Thirdly, there have been a number of metadata standards and metadata initiatives for digital libraries, as I have discussed in Chapters 2, 3, 5 and 8 of this thesis. They are listed again here:

- MARC is the standard for exchange of bibliographic information within the library community and thus provides the key descriptive metadata for use in digital libraries.
- TEI Guidelines represent a standard for encoding textual materials in the academic community.
- Dublin Core Metadata Initiative developed as a basic descriptive metadata for use in a variety of applications.
- EAD provides a standardized encoding format for machine-readable archival finding aids.
- MPEG-7 is an ISO standard for encoding complex multimedia objects.
- SMIL from W3C provides an XML format for encoding structural metadata for multimedia display over the Web.
- OAIS provides a general framework and an information model for digital

preservation.

- OAI defines ways in which descriptive metadata (that is Dublin Core) can be shared between organizations.
- ONIX is a standard for the representation and communication of product information from the book industry.
- OpenURL is a standardized format for transporting bibliographic-type metadata between information services and can be used as a basis for reference linking.
- RDF provides a robust and flexible architecture for supporting metadata on the Internet.

Both Dale [2002] and Calanag et al. [2002] addressed the possibility of how to tie together the existing technologies, and so that they would potentially be able to support the libraries' need to capture and maintain complex and comprehensive metadata regarding objects encoded using those technologies.

XML effort

METS proposes to provide an overall framework within which these metadata schemes and initiatives can be integrated. METS is an initiative designed specifically for use in digital library metadata. It is an XML document format for encoding digital library objects comprised of text, image, audio and video files and is expected to promote interoperability in descriptive metadata, administrative metadata and technical metadata while supporting flexibility in local practice [McDonough and Proffitt, 2001].

METS has three advantages from which I think it has the potential to be promising. The first strength of METS is that it provides a container to accommodate XML data conforming to metadata schemas for various purposes. For common use, the METS schema lists a metadata type for organizations to use in digital library objects such as MARC, TEI Header, Dublin Core, EAD and Library of Congress Audio and Video Technical Metadata and so forth. This encourages the use of the same metadata schema and lowers the cost of local applications while trying to work in conjunction with the METS framework.

The second strength of METS is the link structure within the METS framework. In the Structural Map of METS, objects are modelled as tree structures, for example book with chapters with subchapters. Every node in the tree can be associated with metadata and the link structure registers the links between the nodes. METS allows the metadata to be either embedded directly within its own structure, or held in external files and referenced from within it. Files within the file group section may also be linked to the metadata. All of these complex series of connections within the structure operate by using a variety of linking

elements and XML ID/IDREF attributes that work in association with each other. In other words, the linking mechanism ties every part together in the METS framework, thus those metadata schemas and initiatives I mentioned earlier can work together and integrate with the digital libraries.

The third strength of METS is that it allows institutions to execute programs/behaviours with the information in the METS document and to provide a link to the executable module for that behaviour. This gives METS the powerful ability to support an object-oriented paradigm more completely; that is, a single METS document can identify the behaviours, content, structure, and metadata associated that are needed to make use of the digital library objects. I discussed in Section 2 of Chapter 4 about the theory of the tree structure of an XML document mirroring the structure of an object database and thus performing better while programming.

It is my view that the greatest advantage of METS is that because it is written in XML schema, it inherits the advantages of the XML framework with flexible, modular, extensible, expressive and open characteristics, making it robust and readily interchangeable with other schemas. In this sense, the development of METS contributes to the development of digital libraries with an economic solution.

METS could be seen as a continuation work of MOA2; hence, it is not based on pure theory but has a reasonable basis of practical implementations. This can be regarded as another advantage of METS, that digital libraries need less time to experiment with it and could adopt it in their structures within a shorter timescale. METS is publicly available at the Library of Congress and endorsed by the Digital Library Federation. Institutions can find useful information and tools on the METS Website maintained by the Library of Congress including training sessions, documentation, METS implementation registry, METS extenders (registries and extension schema which are used to extend the METS to incorporate specific types of metadata in a standardized way) and some tools created by the METS community.

Problems with METS

The greatest challenge, according to the main author of METS, Jerome McDonough, is primarily organizational in nature rather than technical [McDonough, 2003]. Since METS was designed to be flexible, it could be adapted to local practices; this means that each institution could define local rules of description, controlled vocabularies and so forth. Meanwhile, METS allows digital libraries to specify constraints that they place on METS including the use of particular extension schemas, rules of description; the arrangement and use of METS elements and attributes for particular classes of documents; specifying the technical characteristics of data files within a METS object. To coordinate the development of readily

available tools for creating and displaying METS objects is another challenge for METS. Digital libraries would need to identify tools for creating and processing METS documents compliant with their particular profiles.

Gartner [2002] also warned of a risk when applying METS in digital libraries. He pointed out that there was a lack of standardization of metadata content. METS only provides a carrier format but the content of the metadata will have been formulated according to different conventions, such as MARC, Dublin Core, EAD or other in-house rules used for describing materials particularly digital objects. Therefore, for METS to be entirely successful there needs to be in addition a greater standardization of content. I think this is the most difficult area to achieve standardization.

Indeed, as Gartner proposed, before the digital library community attains the full benefits of the METS technology, there is still much that could be done via the partnership within the digital library community, including the development of digital library tools, formalizing administrative and descriptive metadata sets for use in the digital libraries and many others. I have discussed many of the initiatives involving collaboration and coordination in the development of digital libraries; perhaps the new technology METS has created another much more crucial challenge that needs the partnership of the whole digital library community, rather than just one part.

8.2.1.2 Conclusion

Michigan uses XML in its image collections and has scheduled the transfer of all text creation to XML from SGML; the American Memory at the Library of Congress is still in SGML, but for completely new projects, XML has been used where SGML had been used in the past. Perseus has already moved from SGML to XML. Thereby, Perseus demonstrates technically the feasibility of this move.

Metadata is the basic element of digital libraries, used to create, describe and manage digital content. Metadata schemas are how digital libraries' data are organized. Metadata schemas play a key role as a quality control mechanism during resource creation and for use in resource discovery. Metadata schemas encoded in XML promise longevity, flexibility and compatibility. Thus it can be seen that the three case studies already implement or are moving towards XML. I conclude therefore that XML has a high impact on metadata.

8.2.2 Access Management

In the electronic environment, digital library access management has emerged as a topic of great interest among many information consuming institutions and information resource providers. A flexible and robust access management service is more than a technical architecture. It also involves a number of difficult issues, including policy and infrastructure considerations, rights metadata in a usable format, deployment of technology in a broad consensus and development of standards among partner constitutions. In Section 1 and 2.7 of Chapter 5, I discussed several technologies relating to rights management in an electronic environment, including DOI infrastructure and Open Digital Rights Language. The research question under this area is:

- Are you working on the area of authentication and authorization, and what is the current approach?

8.2.2.1 Discussion

A wide variety of techniques has been used in access management. Some of them are commonly used in digital library settings. A common form of authentication is the network topological IP-address method. Anyone who has access to a computer with an approved IP address is authenticated. Another standard and widely used method of authentication is remote user registration methods. Off-campus access is becoming increasingly important to many users. Some digital libraries have responded to this by setting up proxy servers. This allows off-campus users to appear to external data sources to be on-campus. Data hosts must provide well-defined terms and conditions of use to assist in the development of licensing and resource-sharing agreements. This is often seen in an open access scenario.

1. Interviews

Perseus

I was told that Perseus had a policy of open access; nevertheless, they begin to move forward with one research area that they have been contemplating, the user interface personalization. But, at the time of interview, Perseus does not keep track of its users. They do not yet create a user profile for any one, so they cannot use authentication or identify who has seen what. All they can do to support a protected online environment is to state explicitly on their Web page relating to copyright of their images: "The Perseus Project does not have permission or resources to redistribute images. No image contained in the Perseus digital library may be reproduced in any form without permission of the copyright holders. We cannot authorize reuse of Perseus images on other WWW sites, even for educational, non-profit use."

Michigan

Academic campuses encounter stricter scenarios on access management. Michigan uses an Oracle database for multi-institution authentication in the PEAK system. This mechanism, along with PLXS, is available as part of other systems or collections that DLPS builds. Michigan has their text collections separately. Collections that are freely available are on one server, and there is very little authentication. These are very widely used, for instance, almost a million searches in MOA every month. Having their restricted materials in a separate server, the first step is IP authentication, and then is individual authorization either through the University Library or through a publisher which hosts materials such as the University of Michigan Press. Also, Michigan has several levels of authentication system to control access to the library system. According to the research interviews, an authentication pilot program exploring systems and tools to reduce risks of anonymous access at public workstations is underway in several campus libraries. This effort is in response to continued concerns on campus relating to network security, aiming to test methods of screening anonymous email sites or requiring authentication for full Internet access.

The Library of Congress

The Library of Congress has no enterprise-wide system that supports both authentication of users and authorization to access a particular application or particular content. In other words, the entire institution at LC does not have any centralized authentication system for its Web-based services. LC thinks it is hard to judge how best to support the authorization needed to handle authorization internally for all the members of Congress and all their staff and for the public. LC protects its files with a firewall, mainly because its users are outside the Library, so there is no system of identification for them. It has to be open access to the public. According to the views of the staff at LC, there is no reason to manage those identities apart from providing access. The purpose of the network firewall is solely to separate internal from external usage of LC systems. LC has rich and varied collections and a heterogeneous user population. Several retrieval issues have been identified but are not yet being addressed, particularly the needs for access control within the search engines. I discussed in Section 1.3 of Chapter 7 how the CORDS project from the Library of Congress has been an experiment on a method of digital copyright registration of dissertations. Also, CORDS has been acting as a digital signature on the documents to check that nothing has been corrupted in transmission. LC uses CORDS in some other applications within its large institution such as, in the case of content, to authenticate that it is the authorized person who has sent it. There is no XML in CORDS because it was developed before XML.

The LC DAVPPP activities will be implemented in the new National Audio-Visual Conservation Center in Culpeper, Virginia, which is scheduled to open in 2005. From the

research interview, I learned that unlike the LC main Website, the Project will place great concern on file integrity, which has the same function as authentication. For managing file integrity, the Project plans to use a metadata file schema registering system at some points within the system. To protect copyright protected digital audiovisual materials, the Project plans to apply a dynamic IP address to support authorized access to certain workstations in the building. There will be password control access on the browser interface for other services. At the same time, the Project has been looking at more robust access management which can be applied as the technology develops.

From the research interview, I learnt that since 2001 some effort has been put into the rights metadata. There is a metadata policy group at LC that has been meeting at intervals of six or nine months, and has produced an internal report highlighting issues on all areas for which the digital library needs metadata. It includes not only traditional cataloguing to help users to find metadata, but also the metadata to support rights, preservation and the manipulation, presentation and use of digital content. That report highlights the need of rights metadata.

2. Analysis and conclusion

Current status and problems in access management

Firstly, Lynch [1998] in his paper states that authentication is the process where a network user establishes a right to an identity; authorization is the process of determining whether an identity is permitted to perform some action, such as accessing a resource. This paper is still CNI's main document on this issue and has not been superseded.

Secondly, DLF sponsored, in 1998, a Digital Library Authentication and Authorization Architecture (DLA3) project, participants including California Digital Library, University of California, Columbia University, JSTOR and OCLC. This project developed an architecture, protocol and operational model for using X.509 digital certificates for authentication and a Lightweight Directory Access Protocol (LDAP) directory service to serve user attributes and to provide restricted access to licensed online materials. In the continuation of the project pilot, XML technology was expected to be included in order to provide improved access, privacy, efficiency, flexibility and richer management information [Millman, 1999]. Since 1998, there has been apparently no progress. The DLF Web page on this topic has not been updated since the year 2000 [DLF, 2005], probably because it relates to the DLF-sponsored pilot in which the University of California (Office of the President), Columbia University, JSTOR and OCLC were involved.

Thirdly, until the release of the W3C Recommendation of XML-Signature Syntax and Processing in February 2002 [XML-Signature Syntax and Processing, 2002], which provides

a generalized architecture for the development of electronic signature applications, there were no standards for the format of the data to be signed, for the format of the signature itself or the format of the electronic documents. This led to the appearance of proprietary solutions and incompatible ad-hoc software applications in the real world environments [Berbecaru et al., 2000].

XML efforts in access management

Firstly, patented digital material protection technology has been growing both in research institutions and commercial sectors since 1990s [Watermarking World, 2005]. Digital fingerprinting and watermarking is another technique that has been developed for the purpose of access management in an electronic environment. This is to place invisible or inaudible marks in certain types of content. In the digital library environment, these technologies are valuable by raising the bar to piracy. Special collections or digital materials with great value would need digital copyright protection to avoid misuse such as by widely redistributing the content without permissions [Watermarking World, 2005].

XML technology is considered vital to the patented digital material protection technology because it provides an independent, open standard supported by major software vendors for the exchange of data. For example, a solution from Digimarc: they use XML in their product because they recognize that it provides the flexibility to operate efficiently and effectively in diverse computing environments, with many different customers, business partners and networks [Digimarc, 2000].

Secondly, the XML-Signature XPath filter is a W3C Recommendation released in 2002 [XML-Signature XPath Filter, 2002]. XML/RDF along with the XML linking ability allows the XML-Signature the ability to select a portion of an XML document to be signed using an XPath transform. It seems to me that this XML approach brings additional and invaluable benefits. Firstly, XML Signature can be implemented with and use many of the same toolkits that a digital library is using for XML applications; thus, there is no need for extra software. Secondly, XML-Signature allows multiple users to apply signatures to sections of XML within a single document, not simply to the whole document.

Digital signatures are created by performing an operation on information such that others can confirm both the identity of the signer, and the fidelity of the information [XML-Signature XPath Filter, 2002]. I note that XML digital signature is becoming important to a growing number of XML applications such as publishing and commercial applications [Bekaert and Van de Sompel, 2005]. I also anticipate that in a digital library environment, digital signature could potentially be used in a scenario such as inter-library loan. Additionally, digital

signatures could be used to indicate a document had not been changed from the original. The use of this in the business world is obvious where it is necessary to know that the terms of a contract have not been tampered with. In the digital library world, one can see instances where scientists, with the benefit of hindsight, might want to change statements that they had made in articles, for instance, to substantiate a claim that they were the first to make a discovery.

Lastly, an encouraging future effort which may have implications for XML is from Shibboleth [Internet2, 2005] which is a project of Internet2/MACE and is being investigated for implementation by the British Athens authentication system [Sankar and Garibyan, 2005]. Shibboleth has been taken into account by some library portal projects in the United States. Gourley [2003] believes that it will be a very long time, if ever, before all of the thousands of content providers implement Shibboleth as an access mechanism. Shibboleth is an open source implementation intended to support inter-institutional sharing of Web resources subject to access controls. The Shibboleth system conforms to the Security Assertion Markup Language (SAML) standard produced by OASIS, which is claimed to be the leading standards body for XML-based technology [Morgan et al., 2004]. As a Web technology, I predict that Shibboleth has the potential to be incorporated into the entire digital library infrastructure, and if the digital library is an XML aware one, it would not be a problem for it to become Shibboleth aware.

8.2.2.2 Conclusion

The digital libraries in the case studies did not find access rights an issue. LC and Perseus are intended to be public access; and Michigan has solved the problem for its institutional users by restricting much of its data to its students by their institutional password access. If access management had in place standards, and these could easily be implemented, no doubt they would be used. However, I can see that the technology is not mature because, for example, as discussed in the previous section, DLF has made no progress since 2000. For this reason, I think digital libraries avoid using access management systems because, though they are desirable, they are not essential.

Access management can be complex, and there is still a lack of a great deal of effort in standardization. There has not been any robust and mature access management infrastructure available as a common management system in the library community. As I have seen, no advances have been made in access management since 1999. My view is that access management decisions certainly will have to be made as an institution level policy. However, access management is not simply a question of developing appropriate policies. In a digital

library environment, the access management system needs to be able to work with copyright and take advantage of the digital environment. It is very likely that in an ideal digital library, users are allowed to search, select, repack, download and at the same time clear the right to use a particular material. I think it would be beneficial for digital libraries to apply XML technology for the universal rights identification standards of content (that is rights metadata) and licensing tools to build a globally electronic copyright-management system where digital libraries could easily share works, rights and information within the global network.

8.2.3 Delivery Systems

My research questions did not focus much on technical issues but rather on general system infrastructure, information retrieval, database issues which I discussed in Chapter 4 and the new technology Web Services as a means for systems interoperability. The research questions included:

- Do you use an in-house developed system or a vendor's system? Is this an XML-aware system? How does the system handle text and non-textual data?
- Do you use the Resource Description Framework (RDF) technology? How is RDF implemented in your system? Does it link to other systems?
- What type of database do you use?
- How many computing personnel are involved in setting up and maintaining the delivery system?

8.2.3.1 Discussion

The underlying architecture in a robust digital library system is often more complicated than expected as digital libraries will be able to move beyond resource discovery and offer services to reach the broadest audiences, that is support a highly interactive environment with knowledge-rich multimedia information resources, allowing a high degree of user involvement and control. The base architecture may be further subdivided to be tailored for various types of information. For example, the extensions for digitized movies will be very different from those for image or audio information as different digital objects need different technical supports. I think flexibility and efficiency could be the biggest design challenges because of the richness and variety of information in the digital library.

1. Interviews

Perseus

Perseus works in a UNIX environment. In the back end database, at compile time, the XML is split into fields and loaded into an open source relational database, PostgreSQL, for fast lookup. For full-text and fielded metadata searching, the XML metadata are indexed by the mg++ search engine. The metadata is stored in XML serialization. In the runtime database, the data are broken into fields to facilitate lookup. Perseus uses another popular open source relational database MySQL for the production servers. Although the metadata is held in relational databases (PostgreSQL or MySQL), the databases merely point to text files and indexes in XML and to images. In other words, the data themselves are stored on disk file systems, not in databases.

As to non-text information, Perseus builds a separate table to handle the peculiarities of Quicktime movies, AutoCAD files and Virtual Reality Modeling Language files. Most of these, like Quicktime files, have a series of tools which are linked one to another. Extensible 3D (X3D), a standard XML format available for 3D graphics, is used to encode the architectural models. The Perseus Atlas interface was written in Perl and uses MapServer for image processing and dynamic delivery of geospatial data which are stored in PostgreSQL. MapServer is an open source Internet GIS application developed by the University of Minnesota in cooperation with NASA and the Minnesota Department of Natural Resources.

To support the interactive Perseus, the system manages multiple DTDs through mappings from elements in the DTD to abstract document structures that facilitate the application of tools in cross citation retrieval, toponym extraction and plotting, automatic hypertext generation, linkages to morphological analysis, generation of maps with a Geographic Information System (GIS) and discovery of word co-occurrence patterns.

From my research, I noticed that one of the characteristics of the Perseus document management system is the use of the abstract bibliographic objects (ABO) approach in the delivery of documents. Each ABO represents a unit of intellectual content in the digital library. The ABO identifier is the key to various metadata tables. It could be text such as a commentary or a physical multimedia object such as a vase. A vase might have more than fifty image files associated with it. This ABO serves as an entry point or reference for specific files related to the object [Smith et al., 2000].

The document display in Perseus is controlled by a template, written mostly in HTML with place-holders for various display elements and portions of the document. The layout specification metadata record for the concrete XML document determines which template to

apply. The decision to use HTML was made mainly because of its universal support. As browser support for XML becomes more robust, Perseus expects to exploit the XSL to produce more versatile interactive XML displays.

RDF had been discussed and experimented with worldwide when I designed the research interview questionnaire. As the nature of Perseus is to be a technology-oriented digital library, it is not surprising to me that Perseus is the digital library initiative among the three that moves most quickly to this cutting edge technology. In Perseus, the RDF is implemented in the following way: the metadata are archived in XML serialization. According to the programmer at Perseus in an exchange of emails subsequent to the interviews, in the runtime database the data are broken into fields to facilitate lookup.

This database structure contains one record per RDF triple (an RDF statement), rather than one record per object described. An example of this is that repeated dc:Creator values do not need special handling. In other words, because of RDF, the Perseus system infrastructure can possibly work efficiently to manage its complex multi-media database. In addition, RDF in Perseus is linked to its OAI harvesting and searching service. Metadata from harvested digital libraries are loaded into RDF format.

The Perseus personnel are structured very differently: every member of the core staff has knowledge of computing. Because their project relies entirely on grants and gets no support from its parent organization, they have to be more careful with their funds. Therefore, they can not afford different groups of people as LC and Michigan do.

Michigan

Michigan Digital Library eXtension Service was developed in a server and application environment. It provides both a search engine (XPAT) and a set of tools for mounting digital library resources. XPAT is SGML and XML aware, and is the heart of the system architecture. XPAT is specially designed to handle large and highly structured documents and metadata such as that found in digital efforts. Support for XPAT is designed around a core set of middleware applications that are distributed freely through DLXS.

Michigan uses "Class" as a middleware in the system architecture when performing information retrieval. The "Class" includes class for text, continuous tone images and associated metadata, bibliographic information and finding aids. In the case of text collection searching, a query (including browsing the collection) submitted by a user will initiate a call to the Text Class middleware. The Text Class middleware will contact the "colldb", a specific instance of Object Name Resolution, to locate the host of and name for the index to the

collection. The Text Class middleware will contact an XPAT index of the SGML- or XML-encoded version of the collection. A request for a display of data then draws the data directly from the SGML- or XML- encoded data in the repositories.

The DLXS is working in the direction of open source software. At the time of interview, they are available for a number of UNIX environments. Versions of XPAT have support strategy for multiple environments, including support for multiple database environments, multiple operation systems. It is available for the Solaris operating environment (Sparc), Linux (Intel), Oracle and soon for MySQL. DLXS middleware typically uses the Apache Web server, and uses an object-oriented approach in Perl. For managing such large amounts of information, an efficient method such as the one Perseus has implemented is necessary. Similar to the approach in Perseus, Michigan does not actually store data in a database. It works through the repository to locate text strings, text structure or other digital objects. The repository is a large file system of ITU TIFF G4 page images, SID files for Image Services, SGML or XML-encoded texts or text collections, or other similar digital objects. Michigan uses the relational database software Oracle to manage information about the collections that they have online, such as the names, the identifiers, what server the collections are on, and the organizations. Michigan was not aware of RDF technology at the time when I interviewed them.

Michigan DLPS has a programmer who works full-time on core programming issues, and system administrators who administer the server. One member of staff who is a programmer is also the content coordinator for the image service, does the programming for creating the services and also works with different groups on campus that have image databases to get them online; two programmers are working on Text Class and finding aids areas. By and large, the programmers are responsible for delivering the materials while librarians are involved in getting materials online. There is another member of staff working on a funded project, who is using text and image classes to deliver their text and images. Additionally, Michigan has one interface specialist who works on all the different features of access, as well as on developing the library Web pages and catalogues.

The Library of Congress

The LC NDLP uses Oracle as its core database system in its information technology infrastructure [National Academy of Sciences, 2000, pp. 204-210]. It is built around a search engine called InQuery, which is a flexible set of tools for application developers. LC takes the features of InQuery and customizes them. It is the underlying engine for most applications at LC including Thomas, which is the public face to the legislative information. The search engine manages indexing the full text of book-length works, and integrates many other library

heterogeneous sources of data including bibliographic records for multimedia materials, non-MARC records and finding aids in a single search [Arms, C.R., 1999].

The main reason why SGML technology is still used is because LC has not found an XML search engine that will do what they want, and there is no immediate thought of switching from their existing search engine. The InQuery is still in use for American Memory along with existing available tools and programs and works very well, so there is no plan to change in the short term. I addressed this issue in Section 5 of Chapter 2 as one of my research conclusions.

According to Arms et al. [1997], LC system architecture roughly consists of four components. First, the user interface performs the user search query. Second, the repositories store and manage digital objects and other information such as databases, Web servers and so forth. Third, the Handle system is a computer system identifying the address of digital objects stored in the repositories. Each digital object has a unique identifier, called a handle. Last, the search system (InQuery) retrieves the digital objects requested from the search query. The interface client services send the handle of the chosen digital objects to the handle system, which returns the address of the repository. The client service chooses the appropriate one and presents the user's browser with the list of digital objects found through the search system (currently through HTML).

Although there was a meeting at LC discussing RDF and its relationship with metadata, librarians at the Library of Congress did not find RDF attractive, as I discuss in Section 4 of Chapter 9.

Although XML is not mainstream technology at LC, I learnt that currently several small XML-related projects are exploring the possibility of XML implementation. *American Women* was a book published by the Library, and the Library selected it and decided to put it up as part of the American Memory collection on the Web. It started out in SGML, using the American Memory SGML DTD, and then converted it to a well-formed XML file using the open source program XS. The XML files were then transformed and presented using XSL technology. The American Women project is entirely XML-based. Furthermore, from my update email on 19 August 2004 with LC staff, I learnt that LC is making some use in other projects of different search engines which are XML based. For instance, Apache Jakarta Lucene is in use for the 'I Hear America Singing' (IHAS) project.

LC has entirely the opposite scenario in computing personnel for systems development. Apart from the group of people who manage the digitization process and organize the content, LC

has different groups of people involved in Web page design and setting up and maintaining the delivery system. In American Memory, four computing people were responsible mainly for programming and indexing whereas design people were focusing on Web page design. Additionally, a technical review group which included all the programmers and key people in the production area, met and reviewed projects on a regular basis. The purpose of the technical review group was to try to ensure that decisions made at every stage were moving ahead in coherent and collaborative ways that the project wanted to move ahead in general rather than trying to do things in a special way for one particular project.

2. Analysis and conclusion

In digital library information architecture, a single work may consist of many parts; several digital objects may be grouped together in order to represent more complex information structures. These can be several chapters that make up one book and may additionally include images that belong to one of those chapters. The digital library information architecture may coexist at the same time with many different software packages, for example, some parts of the digital library may require Excel to view them and the others will require programs such as Word or Acrobat to read them. In other words, digital libraries will require the convergence and interoperability of diverse systems to manage the complex objects. From research interviews, I learnt that the LC DAVPPP planned to connect METS XML to Flexible and Extensible Digital Object and Repository Architecture (FEDORA) [FEDORA, n.d.] as tools for the display of digital objects because the Project thought that FEDORA well supported object behaviour in dealing with massive and complex digital objects.

There are several crucial points that may contribute to the success of a digital library system infrastructure, including the separation of the data and the system, keeping up with the latest technologies, such as XML technology, and standards which contribute to the scalability of the systems. The three systems in the case studies are such cases.

Additionally, my view is that the information infrastructure needs to be developed with a full understanding of the institutional objectives and constraints. Vision and direction must be set clearly before making technological choices because it is difficult to make changes in technology after the investment has been made. The three digital library systems were all developed in-house. In the case of LC, their Information Technology Services did most systems development in-house initially; however, it was rather inefficient to maintain the development as technology advances so rapidly. Increasingly, LC started to use third-party software rather than developing it in-house. Examples of such are the core technology, search engine Inquery, which is a commercial solution and an ad hoc XML-aware delivery system, including a tool supporting XSLT particularly for METS presentation in the LC DAVPPP

which is developed by a contractor. Also, I learnt from the research interviews that the mg++ search engine in Perseus is available free of charge from Ian Witten's group at the University of Waikato, New Zealand; XPAT in Michigan was developed based on the search engine Open Text (Pat), which was originally developed in the University of Waterloo, Canada.

I pointed out in the conclusion of Chapter 2 the problem of the lack of availability of XML tools such as general purpose text editors for building digital libraries. In the case of search engines, Perseus looked for an XML search engine and found one in the open source community as mentioned above; Michigan developed their own SGML and XML search engine, while LC is still using SGML because they cannot find a suitable XML search engine to support their work. Although SGML and XML share a common syntax, Flynn [1998] thought it worthwhile to point out to users to consider for any tool if it works with XML as well as regular SGML. SGML-conformant tools for editing and syntax checking may be compatible with XML; others such as search engines may be compatible after modifications. XPAT in Michigan is one such case which has been modified for XML. XML was created to overcome the limitations of SGML for applications on the World Wide Web, so many new tools for XML will have to be created to support the XML family of specifications as they become mature with time. It is fair to say that more tools are being developed in the XML open source community than in the XML commercial world. One reason may be that commercial vendors would not invest fully in still-developing technologies. The other reason may be that, as Banerjee [2002] pointed out, since XML is still new, people are concentrating on implementing or experimenting with XML for their own purposes. Indeed, many XML open source tools are being developed by research institutions and library institutions themselves for their digital library applications, for example, Open Source Digital Library System (OSDLS) [OSDLS, 1999] and Open Source Systems for Libraries (OSS4LIB) [OSS4LIB, 2005]. Michigan, as mentioned above, has been developing XPAT. Although XML itself reached W3C Recommendation status in 1998, XML editorial tools for the creation of XML documents are not yet easy to use when compared with the HTML editorial tools, for which users do not necessarily need knowledge of HTML. Because no such easy-to-use XML tools are available, the Medlane project planned to develop editorial tools for librarians to edit their bibliographic and authority records without having to know XML [Clarke, 2002]. SGML is solely used in government, industry and academic research projects, and never reached the same level of popularity as HTML has done. It is reasonable to assume that commercial vendors, research institutions, and library institutions will continue to develop tools that take advantage of the trend of XML technology in Web applications.

In digital libraries, the users are looking for intellectual works by searching through raw metadata, and the software achieves this by tools and systems that can take advantage of the

metadata. Each digital library system provides the options for configuring indexes to support metadata discovery, finding and collocating.

Problems in delivery system

Perseus raises issues of using advanced technology in the overall quality of the information resources and navigation in the corpus of a digital library with widely varying encodings and markup practice. Highly structured markup language, first SGML and now XML, have been contributing largely to improve the extraction of structural and descriptive metadata from documents and to deliver document fragments on demand. Furthermore, the richness of implicit and explicit links between information resources has been helping to meet the challenges of automatically generating hypertexts in electronic media.

In the emerging information-rich society, there will be an increasing demand for high quality, enriched digital multimedia content some of which will be satisfied by the digital library. Information and communication systems in the digital library will be able to create and deliver multimedia content to meet the educational needs of different levels. To this end, my view is that interoperability within digital libraries networking is not simply a matter of providing coherence among digital object repositories, but also technical interoperability that supports a broad range of inter-repository protocols, distributed search protocols and technologies including the ability to search across heterogeneous databases. For working towards greater interoperability, the availability of markup language can be influential in enhancing access to materials that are inaccessible through descriptions because their descriptions are non-existent or of low quality. It is not easy to maintain the quality of data identified by the markup in such a way to facilitate the exchange of information between systems and the migration of data to new systems.

XML efforts in delivery systems

Firstly, I indicated in this section that in the real world, many digital libraries are built on a variety of loosely coupled commercial, open-source and home-developed components. Digital libraries will need different Web applications to work together and provide integrated services to the users. Interoperability of these applications into a consistent system is a challenge. The new XML technology Web Services may provide a more flexible environment for a digital library infrastructure. Web Services allow applications to communicate with each other by sending XML-encoded messages over standard Web network protocols. In a digital library environment, I anticipate that this can be served to applications such as on-line document ordering in an interlibrary loan service or to perform broadcast searching of physically separate databases held in different institutions which overcome several of the limitations found in OAI.

The Web Services are a common and universally accepted grammar for integration. They use XML to describe services through the Web Services Description Language (WSDL). Also, they use SOAP (a W3C XML-based protocol which supports Remote Procedure Call in a decentralized, distributed computing environment) to pass messages between services and the client applications which use them and HTTP as the transport protocol. The Universal Description, Discovery and Integration (UDDI) is a special Web Service which allow users and applications to locate required Web Services. The UDDI, WSDL and SOAP form the Service Oriented Architecture (SOA), which is expected to leverage existing investments and work over the Internet with features of flexibility, productivity and cost savings [Haas, 2003].

Web Services have been identified as one of the Top Technology Trends in 2003 and 2004 Midwinter and Annual Conference by the Library and Information Technology Association (LITA), a division of the American Library Association (ALA) [ALA, 2003]. Web Services technology is increasingly attracting the attention of the library community. The Washington Research Library Consortium has used Web Services as a methodology in integrating digital library systems [Gourley, 2002]. The LEADERS project being undertaken at SLAIS, UCL is another example. The project uses WSDL to describe services which are utilized by applications; the SOAP XML messages are generated to carry messages between applications and toolkits. The same application can be used to access different resources served up by the Web Services [Turner, 2003]. Denmark's Electronic Research Library (DEF) has been exploring by a number of projects building a digital library with XML Web Services. The open source FEDORA system (METS based) is being implemented as a digital objects repository system [DEF, 2005]. Furthermore, the Library of Congress NDMSO is developing MARC XML Web Services as part of the framework for working with MARC data in an XML environment [NDMSO, 2005]. Felstead [2004] provided evidence of the practical application of Web Services which can contribute to interoperability between integrated library management systems and external systems. From my update emails, I learnt that Perseus had recognized the value of Web Services and had implemented them in their infrastructure [Mimno, 2004]. These are examples of the fact that the XML technology has potential in playing a crucial role in integrating Web-based application systems.

Web Services are still very new, but are likely to provide the foundation for many new models of cooperation and integration among Internet-based applications. It seems to me that they have been discussed more in an electronic business environment. The digital library is an Internet-based implementation. I estimate that Web Services may possibly provide digital libraries with a more scalable, cost-effective, integrated and open base to develop, which gives them more chance to succeed.

Secondly, in an area that concerns both traditional libraries in their integrated library systems and digital libraries, standards are being developed for the interchange of circulation information, which will ultimately enable material belonging to one library to be loaned to a member of another library and will pass between systems all data relating to the transaction, such as date of loan, period of loan, any fines incurred for non-return of material and the like. It was agreed in 2000, early on in the development of the NISO Circulation Interchange Protocol standard (NCIP) [NISO, 2002] that it would use XML for encoding the messages which will be transferred between systems.

Thirdly, the three case study libraries had been very interested in the XSL stylesheet language as a presentation. As demonstrated in the Ching project in Section 7 of Chapter 4, XSL is used to format XML information for display, and it can also be very useful when managing data interchange between different computer systems in Web applications like digital libraries. As XSL is built with presentation flexibility, a digital library can use an XML document to create a number of stylesheets for different scenarios such as new books lists and an interlibrary loan form.

I noticed that the three case study libraries have demonstrated that the role of relational databases is important as well as the best choice for storing the data. On the other hand, they implement the concept of object-oriented paradigms which indeed promotes the maximum efficiency while the data remain in a standard format with the power of XML or SGML. I discussed in Section 3.3 of Chapter 4 that the relational database vendors are introducing more object-oriented concepts into the relational database which is having the effect of strengthening the position of XML.

In general, I see the three digital library initiatives have been developed within an open architecture to comply with international standards and formats in order to promote interoperability with other systems and browsers. The content markup has been implemented in SGML and XML based on TEI with extensions which include pointers to text and non-text materials. As a result, a high degree of granularity has been achieved which has facilitated the development of a variety of profile tools which fully exploit the multiformat resources.

Interoperability between heterogeneous systems of information resources and services is becoming of more concern when building digital libraries because a flexible and efficient digital library infrastructure requires interoperability. With this in place, access and retrieval are enhanced. The adoption of a set of markup features smaller than SGML, that is XML, is expected to be able to integrate fully with Web applications. Perseus has demonstrated the potential of XML in interoperability. As we have seen in this section on XML applied to

delivery systems, my conclusion is that XML has a high impact on system interoperability in digital library development.

8.2.3.2 Conclusion

Increasingly, the Internet has been transforming from an information provider to a service provider. To achieve seamless access to the large-scale repository as the basis for digital library services, several aspects needed to be considered. I conclude that the first and fundamental aspect is a widely accepted metadata standard which allows for enhancing information retrieval. The metadata will need to be good and strong in order to enable the metadata records to exist independently over time through successive generations of computer hardware and software or to move to entirely new delivery systems. The second aspect is to encourage the use of standards in an open environment which facilitate the exchange of information between systems, thus promoting interoperability. The third aspect is to develop a mechanism with tools to enable the integration of various formats of resources to deliver smoothly the data to users. XML technology is not tied to a particular vendor or operating system; a digital library has many choices to select the tools that provide the maximum infrastructure and functionality. Through my research and research interviews, I concluded that XML will have high impact in metadata and interoperability in the digital library development.

Implementing XML technology in a library setting in the real world indicates that digital library initiatives have recognized the benefits of XML. There are still a great number of research areas to be explored while applying XML in the digital libraries in terms of what the digital libraries want to accomplish with it. I anticipate that the power of XML makes it possible to integrate Web services and resources, whereby librarians will discover the value of XML and wish to use it to improve services and processes in digital libraries.

My three case study libraries had been monitoring and exploring the importance of advancing the use of XML in the real world in response to changes in the technical environment and the expectations of the users. From comments made during research interviews, I conclude that they will apply more XML technology when support becomes more widespread.

Chapter 9

Maintaining the Digital libraries

This Chapter discusses the management and organizational aspects of the digital libraries in the case studies, including the audiences for which they are intended, illustrated by comparing their good and bad practices. The Chapter concludes by mentioning the main innovations that the libraries in the case studies have made.

9.1 Managerial Aspects

The development of electronic information brings new management challenges to libraries. New tasks have emerged calling for different skills and new input of resources. These involve new information strategies including planning, introducing, monitoring and improving electronic services. Also, there is a growing demand for staff to be associated with the multiple skills related to the creation, acquisition, recording and management of data available in new or expanded areas of activities.

Markup has become the key to the efficient storage, location, retrieval, analysis and evaluation of information. People with knowledge of this kind of information, either computing personnel or skilled librarians, need to be recruited, or existing staff will need to have more training, particularly in XML.

Moving to digital library initiatives, the costs of maintaining and development will shift to equipment and digital conversion; certainly, the expenditure will go on personnel, as human capital is a key resource of this kind of activity.

To some extent, Web statistics are useful for digital libraries to acquire information on the actual load on the server. This is valuable for diagnostics and planning, and for monitoring user behaviour for future development. The investigation of these issues was undertaken by asking the following questions:

Staff infrastructure

- How has staff structure changed due to digital library development, especially XML-based? What are the implications for staff recruitment, retention and development?
- What kind of people do you need to hire to do the work, particularly tasks related to XML technologies?
- Do you provide staff training? What skills do they need?

Cost centre structure

- Where is the money spent? Which areas are the significant parts of the cost?
- What is the cost of development and maintenance?

Maintenance and future development

- What is the growth rate? Could you please provide statistics?

9.1.1 Discussion

9.1.1.1 Staff Infrastructure

Technological advances have brought great changes in library and information services. The changes are set to continue as libraries are expected more than ever to become fast moving, reorganizing and innovative [Arms, W.Y., 2005]. To be able to incorporate well into the roles, I discovered from the research interviews reported in the following sections that librarians require a wide range of new and enhanced hard skills, such as technical knowledge, and soft skills, such as vision for the future.

1. Interviews

Perseus

The view of Perseus was that using digital resources to do research was beginning to have a great impact in the Humanities. Basically, Perseus looked for people in the Humanities with little computing experience and trained them to do research and to create electronic resources. I was told that this had one major impact on the way that Perseus structured the work; that is, wherever possible Perseus did its work in-house, preferring to use its own resources to train and support young scholars in the Humanities rather than supporting outside professional contractors.

Michigan

Staffing in Michigan has grown with its accomplishments. Initial staffing was set at the levels that are necessary to provide a baseline of commitment to all areas, with growth expected for new formats and for extending DLPS commitment to issues such as cross-collection and

cross-format integration. At the time of my research visit, DLPS had approximately 24 full-time equivalent staff; a small part of them were short-term-funded people related to a specific grant. I learnt that Michigan provides limited staff training to all staff but focuses on digitization and encoding skills to staff that are in charge. It seemed to me during the interview that Michigan staff have the most skilled digitization ability among the three case study libraries.

Staff in Michigan are grouped into two areas of work. The Digitization group focuses on methods and formats and is responsible for production-level creation and conversion of digital library resources supported with markup technology; the Information Retrieval and Architecture group build the digital library infrastructure to ensure a smooth delivery system. This group also works closely with the Digitization group to ensure the extremely high volume, high quality digitization operation. Many programmers within this group have areas of specialization including XML and SGML. This ensures a high degree of technical and format understanding in building online systems. The infrastructure work involves most areas of DLPS operations such as interface specialist, data loading and technical support for DLPS staff. During the interviews, I was impressed that DLPS had been working closely with the university librarians, exchanging information across units or within the Library as a whole. For example, library cataloguers are acting as DLPS's metadata specialists. Whenever necessary, DLPS would consult University librarians about naming collections or metadata mapping.

At the beginning, Michigan had more staff with a Humanities background, but now they are getting more staff who have a computing background with interest in the Humanities. Among them, some people are librarians who have just got involved working in computing; some have a really complete programming background; and some a Humanities degree. In the future, Michigan is planning to recruit more people with computing expertise and experience in digital library evaluation, especially relating to user needs, to cope with the difficult work in developing an increasingly complex digital library information infrastructure.

The Library of Congress

LC highlighted the need for changing the skills and emphases of the organizational changes in the Library. Skills associated with particular data types such as XML and SGML have become increasing important to staff.

To support the NDLP, LC developed a well-organized staff infrastructure as follows [Campbell, 1995].

- **Curatorial Staff:** Staff assigned to the curatorial divisions prepare and process materials to be digitized. Curatorial staff also performed on-site digitization of materials that include rare and fragile items such as early drafts of the Declaration of Independence and the Gettysburg Address.
- **Core Staff:** NDLP core staff worked with the Library's divisions to prepare and describe the collections, verifying the status of copyright and seeking permission for use of the materials when appropriate, and digitizing the materials and verifying that they adhere to the Library of Congress's standards of quality. Digital conversion specialists in the central office provided project coordination and technical oversight. The more experienced specialists oversaw collection development and production, serving as team leaders and as brokers among the division and automation staff and contractors.
- **Infrastructure Staff:** Infrastructure staff were primarily information systems experts who built and maintained the automated systems that stored and provided access to the digital collections.
- **Educational Services Staff:** The educational services staff focused on educational outreach for the use of the historical collections by school children between the age of 5 and 18. They researched user needs, talked to the education communities, evaluated technologies for delivery of digitized materials, coordinated collection selection and developed and supervised contracts.

I learnt that a great deal of effort had been put into formulating new job descriptions as NDLP was a vast and new task. LC found that personnel who could bridge the gap between traditional librarianship and technical skills such as those who could build delivery systems were highly valuable and in great demand.

At the peak time, NDPL employed as many as one hundred people with various professional skills related to the needs of building a large digital library. LC provides several kinds of staff training packages both on-site and off-site and has sent many staff to training courses.

It is worth noting that LC staff have been actively involved in activities related to XML technology. For example, when the author was doing the research interview in LC in September 2002, there was a workshop being held there by a commercial institution on the subject of XML-based TopicMap technology, which can be used in information retrieval in a digital library. LC core staff have been interested to know more about XML because they knew that they needed to prepare and know the technology well to decide how XML could be helpful in the digital library development when the technology is ready for them.

2. Analysis and conclusion

Technical challenges in the digital library have, in general, already been recognized, while the non-technical challenges encountered by digital library developers have been proving to be more enduring, complex and profound. The digital library research community is increasingly concerned with the need to base the design of digital libraries on the work of the community they support [Star and Bishop, 1996]. The social impact and influence of digital libraries have raised active discussion on many occasions over recent years, such as at conferences on digital libraries [ACM, n.d.]. The challenges fall into several aspects across the digital libraries spectrum, including organizational co-ordination between a number of partners, licensing, and copyright concerns regarding online materials, user studies and so forth.

Pinfield [2001] examined the new roles and new responsibilities that librarians are now involved in as new era digital librarians. He thought librarians would act as:

- Multi-media users: librarians would be able to feel comfortable with a wide range of formats.
- Intermediaries: with a good knowledge of sources and user requirements.
- Enablers: proactively connecting users with information they require.
- Communicators: formally and informally liaising with users.
- Project managers: leading on development projects to enhance the service.
- Trainers/educators: taking on a formal role to teach information skills and information literacy.
- Evaluators: sifting free and purchased resources on behalf of users.
- Metadata producers: librarians would be able to create records of information sources in a variety of schemas.
- Team players: librarians would be able to work with colleagues in library and information technology services and with academics.
- Negotiators: librarians would be able to deal with publishers and suppliers.
- Innovators: librarians would not just follow the routine but also look at new ways to deliver the service.
- Fund-raisers: librarians would be able to work for greater income from the institution and beyond.

My view is that in Pinfield's list, innovation would be the key requirement for librarians in the fast-changing environment of the profession. Institutions are not looking for people doing routine work but for those who have knowledge of future trends in their professions. Therefore, my view is that librarians in general, following the lead of those in the Library of Congress,

are wise to see the future in XML. The results of my investigations of real world library posts discussed in the following paragraphs support my vision of this.

The Association for Library Collections and Technical Services (ALCTS), a division of the American Library Association, Continuing Education Task Force undertook a survey as a response to Action Item 5.3 of the Library of Congress's action plan "Bibliographic Control of Web Resources". This survey discovered that 73% of respondents thought that cataloguers needed knowledge of XML today and would for the near future at least [ALCTS Continuing Education Task Force (Action Item 5.3), 2003]. This could be that cataloguers in the United States knew of the efforts relating to XML MARC from LC NDMSO; also, it could be that cataloguers knew that the systems they were currently working with were XML based library systems.

In order to examine the theory of the need for XML knowledge as against the perceived need for XML as reflected in library job advertisements, at the beginning of September 2004 (since September is one of the peak months in the year for advertising jobs), I did an investigation using Web job listings in the United Kingdom, the United States and Taiwan. I targeted academic librarian jobs and computing jobs in the academic sector. Furthermore, I also investigated whether in those three countries XML was part of the curricula in library schools and if there were reports or activities related to XML from the library associations in the three countries. Below are my findings and analysis.

In the United Kingdom, between July and the beginning of September of that year, there were none out of 31 for professional or managerial jobs listed in jobs.ac.uk that needed XML during that period; 2 out of 74 put XML required in the job description for professional or managerial computing jobs; 3 out of 14 need XML for computing technician jobs.

In the United States, there were 3 out of 151 librarian jobs listed in the Association of College & Research Libraries across the country that stated that they needed XML in the time period 25 May to 1 September. 15 out of 153 academic computing and information jobs listed in EDUCAUSE (a US association which promotes the intelligent use of information technology in higher education) job listing needed XML in the time period 2 July to 7 September 2004.

In Taiwan, between 9 January and 3 September, none of the 25 librarian jobs listed in the Library Association of China (LAC) job listing needed XML, but LAC gave 3 training courses in XML in 2004 under Digital Archive and E-learning schemes; none out of 7 computing jobs listed in the National Youth Commission Website (a government job Website)

needed XML in the academic sector during the time period 17 December 2003 to 3 September 2004.

In the library sector, most of the jobs which required XML are for posts in areas such as digital library projects, electronic services and metadata while in the academic computing sector, knowledge of XML was stated as being needed more for computing technicians than for the managerial level. However, it is interesting to note that for managerial level library jobs, institutions are looking for candidates who are innovative, creative and able to provide strategic direction and vision for the libraries; they must be able to provide both traditional and innovative library resources and services, possess an informed vision of the library in the 21st century, knowledge of new trends and emerging technologies in information services, and be capable of identifying future needs and directions in an electronic environment. For instance, for the post of Bibliographic Systems Manager, candidates must have substantial innovation and creativity to identify new ways to use bibliographic services to delivery library resources to users' desktops; for a job as Head of Cataloguing and Acquisitions, they must have knowledge of current cataloguing standards for all formats and awareness of emerging trends and technologies in technical processing; for professional library jobs, institutions are looking for a knowledge of IT issues, ability to learn and apply new technologies to improve work operations, ability to handle multiple responsibilities in a changing environment, non-traditional thinking with regard to library collections and services, ability to meet professional standards and competencies and so forth. For instance, for a post of Digital Projects Librarian, they must be familiar with structured markup and knowledge of metadata standards and best practices in digital projects; for a job of Reference Librarian, they must be aware of current trends and emerging technologies in the delivery of reference services; for a post of University Librarian, they must possess a wide understanding of the current technological information environment and an innovative vision for the library as a partner in developing approaches towards access to information; for an Information Systems Librarian, they must have knowledge of metadata standards and best practices. These words do not explicitly state what experience would qualify a candidate but, in a sense, they could be XML-related technologies. Incidentally, job requirements in Taiwan are customarily relatively short.

I also noticed that in the library job sector those jobs relating to automation or digital libraries did not mention XML though some mentioned knowledge of markup language and best practices in digital projects. The lack of mention of XML may be due to a lack of understanding of it by library managers, or due to the fact that knowledge of best practices as stated in the job descriptions covers more than one technology and XML is intended to be one of them.

It is interesting to note that there is a relatively small library job market in Taiwan. Librarianship in Taiwan awards a bachelor's degree to students who have spent four years studying librarianship and information technologies. Students spend two years for a master's degree for the study of advanced librarianship subjects. Librarianship students find jobs in various fields as the line is blurred between librarianship and information technology.

Library schools have not recognized the need for XML skills. According to the course description listed in the library schools, in the United Kingdom, only one (University College London) out of 8 library schools provided an XML course; in the United States, two (University at Albany, SUNY and University of California, Los Angeles) out of 50 library schools provided an XML course; in Taiwan, one (Hsurn Chuang University in Taiwan) out of 10 library schools provides an XML course. On the other hand, the concept of XML has been introduced in several subject courses, for example in Electronic Publishing (Sheffield University and University of Illinois at Urbana-Champaign), in Document Engineering (University of California, Berkeley and University of Illinois at Urbana-Champaign), in Technologies in Web Content Management (Syracuse University), in Information Organization and Access (University of Illinois at Urbana-Champaign), and in Information Technology Tools and Applications (San Jose State University), in Access Systems for Archival Materials (University of Michigan) and in Taxonomy, Classification, and Metadata (University of Washington); in library school workshops (Kent State University and Tamkang University in Taiwan); or has been experimented with as faculty research projects (The University of Strathclyde, University of Illinois at Urbana-Champaign, Tamkang University and many others).

Overall, XML appears in more course descriptions at the University of Illinois than anywhere else. This is probably because teaching staff there have more knowledge of XML. I noticed that in Taiwan there are at least two doctoral-level computing specialists in each library school, giving courses in database management and programming related courses; therefore, there would be no problem for the departments to provide XML-related courses.

In the United Kingdom, the Chartered Institute of Library and Information Professionals (CILIP) has been announcing more news and training workshops on library and information technology in which XML-related initiatives such as markup language, schema, RDF are part of the topics [CILIP, 2004]. CILIP also published Hey's [2004] article on the vital role of academic librarians as metadata experts and digital curators in the digital age; he mentioned that librarians should be well versed in current new technologies such as XML. He felt it is important to mention in this context where the readership consists of librarians at all levels that senior librarians represented by the Consortium of University Research Libraries and the

Research Support Libraries Group had shown little vision in this direction though he conceded that JISC was putting effort into these training tasks.

In the United States, since 2000 the Library and Information Technology Association at ALA has selected XML related initiatives such as MARC XML as annual top technology trends [ALA, 2003]. The Library Association of China (LAC) in Taiwan provides irregular XML training courses. This could be because there is a five-year nation-wide National Digital Archive programme and the training sessions are designed to support that programme.

To conclude my investigation, although XML has not been recognized as a core skill in library jobs or as part of the core programme in library schools, nevertheless, library associations have identified XML as an important technology trend that needs to be monitored carefully. Indeed, as Felstead [2004] pointed out, XML was already being extensively used in a substantial number of library systems, though the use of XML was transparent to the librarians using the systems. I suggest that it would be advantageous for librarians to have knowledge of XML even if they do not work directly with XML; and library schools could provide selective courses on XML.

Very often, digital libraries are part of larger organizations; therefore, there will be a need to review the fundamental organizational structure of how staff are obtained, trained and retained in order to carry out the wide range of library services in the digital age. My view is that staff retention is important and it is necessary to have a good career structure in place, which is going to be difficult in most environments where academic digital libraries are being developed. Young staff will want to see a career path ahead of them with the opportunity to earn more money as time goes on. A digital library where the only benefit for the workers is pride in their work will not help in their retention. Being in a pioneering field with skills in a tool like XML which is used in the commercial world will mean that young workers will be in a good position to seek posts in the commercial world rather than in the academic one. Small organizations will find it difficult to provide a career path, and large organizations like the Library of Congress will also have the problem that employees working on digital libraries may find career promotion within the organization, but outside digital libraries, to the detriment of digital library development.

Inserting markup in a text is an act of interpretation [Hockey, 2000, Chapter 3]; this will need domain specific specialists to deal with the creation of markup in the documents. As Perseus is a Humanities-based digital library, there is a rising demand for corpus editors who combine technical and traditional humanistic expertise to accomplish particular domain-specific goals. The main job of corpus editors is expected to be to establish a reasonable level of precision

that balances scholarly standards, and also to document clearly the level of precision employed to meet users' expectations when reading documents in a corpus [Crane and Rydberg-Cox, 2000]. Crane remarked when interviewed that Perseus had been fortunate in being able to recruit true corpus editors who successfully created the Greek and Latin texts that tied together scalable methods of tagging and a specialized knowledge of classical languages and literature.

Perseus has a very different staff training model from the other two case study libraries. Perseus exposes young scholars in the Humanities to the new technology in a dynamic learning environment. Core staff (about three to five people at a time) are expected to learn XML as a basis. Some staff focus on the most demanding and rigorous software engineering; others may concentrate on the humanistic content more directly. By and large, most of the Perseus personnel are computing specialists to one degree or another, and everybody wears many hats; that is, Perseus staff are able to manage the whole workflow production from content creation to programming and putting the content on the screen.

I noticed that the Perseus' managerial staff infrastructure is heavily constrained by the budget. Compared to the other two case study libraries, the Perseus staffing group is small but dynamic and efficient in terms of the size of the collection they have created. My view is that this can be regarded as one of the outstanding features found in the Perseus environment. Staff in Perseus do computing and creation and they thought it was an advantage for them because there was no communication gap, and hence greater efficiency. On the other hand, Perseus staff are not librarians and do not have support from the University librarians when dealing with the library area of work. For example, cataloguing librarians could be the best metadata people to be consulted. I suggest that this could be a disadvantage for Perseus. If Perseus was not just a departmental activity but was formally integrated into the university library, this problem could be resolved. I have found that it is quite common for digital library projects in institutions to exist completely apart from the traditional library to the detriment of both parties, though they do not admit it explicitly. For example, in the Website of the Library of Waikato University in New Zealand [University of Waikato, 2005], there is no hint that the New Zealand Digital Library is produced in the same University's Department of Computer Science.

Both Michigan and LC have strong support from their libraries with overlap in staff. Perseus did not have the problems of organizational change as it was created as a new service but it has no contact with the University Library. From the three cases I studied, I noticed that there is a substantial difference between institutions in staff development schemes, as they are not equally positioned to supply training opportunities. Mostly, this is subject to budgetary

restriction and lack of staff time. Additionally, the reorganization of the library staff into a team-based learning organization would require libraries to think about how, by whom, and in what combination certain tasks can be done in order to fulfil the tasks successfully. LC and Michigan are examples. Furthermore, I also noticed that librarians in Michigan and LC with many years of experience act as core staff, supervising staff specialized in different areas and controlling digital library processing and future development. The core staff are attending more XML courses, evaluating the possibility of implementing XML.

To develop an ongoing and innovative staffing approach well, there has been a positive encouragement to publish, offer conference papers or become engaged in national and international activities. Perseus was a good instance of this case; my view is that this could be regarded as one of the successes of the Perseus marketing strategy.

It is interesting to know that my case studies needed more computing staff as time went on. In the case of adopting RDF technology in digital libraries, I see that Perseus staff are computing people and collection creators; therefore, they could decide to use RDF on their own initiative. On the other hand, LC core staff coordinated and made decisions on things but they were not computing people; although they learnt from meetings about the advantages of RDF, they were not interested. Michigan librarians were not familiar with RDF or never heard about the technology, not to mention use it. The RDF case indicates that librarians are in charge of the digital library development but as they lack computing knowledge, they may not be able to take advantage of the cutting edge technologies in the first place. This again highlights the importance of the librarians' skills in the digital age as indicated by Pinfield.

XML is a new technology for the Web and it could play its part in every library operation. Librarians in the future will play a mediating role between the computing professions and the users. Knowledge of current standards and newly emerging technology trends such as XML will be a beneficial skill for librarians while looking for jobs either in the library sector or information-related sectors. Library schools could therefore contribute to the acquisition of this professional knowledge by covering these subject materials in their curriculum. As Hey [2004] suggested, librarians should be aware of the technological trends and be prepared, in order to compete and survive in the ever-changing environment.

9.1.1.2 Cost Structure

1. Interviews

Perseus

Perseus always tries to recruit full-time staff who can concentrate on the work, but because of

budgetary restrictions, they can retain only three to five research staff at a time. Travel is also essential for Perseus as meeting the right people in conferences or in the foundations for fund raising is a big issue for Perseus. Perseus spent little money on software. They are entirely in the open source community, and pay no significant line item in software. They spend $500,000 on Perseus' running costs a year.

Michigan

Apart from high costs in staff, Michigan spent a fair amount of money on outsourcing the keyboarding of the markup, but they generate revenue as well. As Michigan DLXS is a cost-recovery operation, fees earned from DLXS software are to cover development, maintenance and support activities associated with DLXS. I was not able to obtain the amount of the full cost of the entire digital library. The budget information available from my research interviews was $600,000 for the academic year 2003 for the Information Retrieval (IR) units alone.

The Library of Congress

The Library of Congress NDLP spent $60 million in five years on its entire operation, and new money has to be raised to support more digitization. LC had a unique privilege that brought NDLP to the public; that is, the regular funding streams from the combined support of the United States Congress, the Executive Branch, and America's entrepreneurial and philanthropic leadership. Also, I was told from research interviews that none of the staff in the Library of Congress has real access to the cost information as the cost structure is not necessarily made clear to the project.

2. Analysis and conclusion

In a digital library, the costs are mainly staffing as staff play a key role in developing and maintaining digital services. Compared with $500,000 running costs per year in Perseus, LC NDLP spent $60 million in five years (1995-2000) and Michigan continues to spend $600,000 only in the Information Retrieval department. The entire expenditure will clearly be substantially greater. Below is a table, for comparative purposes, of data for the three libraries which was gleaned from the research interviews.

Digital library	PDL	Michigan DLS	LC NDLP
Annual cost	$500,000	$600,000	$12 million

(Note: in the case of Michigan, the cost covers the Information Retrieval Unit only)

It is interesting to note that, in the interviews, information on costs was provided only in the interview on Perseus; LC avoided detailed discussion of the breakdown of the expenditure;

Michigan was unable to provide full cost structure. I think Perseus was different because in the interview on Perseus, the creator himself was proud to give evidence on how they built the Perseus digital library with little cost. In the next section, the statistics on the requested figures give evidence of this. Furthermore, I noted the small space that Perseus occupies when compared with the relatively large space of the projects in Michigan and LC. On the other hand, there are advantages of being in an enclosed space: research staff work together in the same room and can easily exchange their thoughts.

XML-based Web services are leading us towards a new application framework which builds upon open source software, standards and interoperability. I see open source software playing a significant role in helping digital libraries to start up. Not all open source projects are the same. Some are too small, some have died from lack of continued support, and some suffer from quality problems. My view is that Perseus is a successful open source project. Michigan develops open source software and charges for other software as well. They license software to other institutions, identifying a clear revenue stream associated with the software technology. LC places less reliance on open source as they are privileged to have richer capital support.

9.1.1.3 Statistics

1. Interviews

The three case studies were asked in the research interviews about statistics. Perseus pointed me to the Website, from where I was able to extract the data. Michigan and the Library of Congress agreed to send me the statistics I required later by email.

2. Analysis and conclusion

Statistics relating to user needs and user behaviour can be acquired via systematic monitoring of the usage of resources. Usage statistics are now available for many online resource suppliers, including digital libraries, and to some extent they provide reliable performance indicators. My view is that analysing the statistics could have the following advantages. First, statistics help digital libraries determine the true impact of Web servers. By measuring the popularity of materials as well as identifying the frequency of the sites that access the servers, usage statistics provide valuable marketing information for a digital library. Furthermore, by determining the paths users follow and analyzing the sites from which they come, usage statistics help to find out which outside sites are most important to a digital library. In addition, organizations, whether commercial or educational or nonprofit, would need solid numbers to make a case for raising funds for their work, so the statistics can be helpful.

Many authorities suggest we must treat the statistics with caution, especially when trying to draw comparisons among institutions, as numerous factors would affect the hits [ITMT, 2001], for example what transactions are recorded, whether the statistics record logins, accesses and searches, how documents are structured, how cacheing systems operate and so forth. Incidentally, the COUNTER organization (Counting Online Usage of NeTworked Electronic Resources) has developed guidelines for recording statistics relating to online accesses, and the organization also developed in July 2004 an XML DTD for this purpose [COUNTER, 2004]. Activities in this area are ongoing [COUNTER, 2005]. It is the first initiative to set practical international standards for the recording and reporting of vendor-generated usage statistics [Shepherd, 2004]. But these guidelines were not available to my case studies when they were under development and have not been incorporated since. In any case, statistics could be more useful for commercial systems where libraries need data to decide whether it is worthwhile to continue to subscribe to a digital library or otherwise. Accordingly, I detail in the following paragraphs how the statistics were recorded in the three case studies and try to make comparisons on the same basis.

Figures 12, 13 and 14 summarize WWW total pages requested for Perseus, Michigan and LC. Figure 15 is a comparison of the three libraries. These graphs were prepared by the author using the data available from the Websites of the project, and in the case of LC and Michigan, the data were provided on request. The statistics covered three fiscal years 2000-2002 and were analyzed monthly.

The Perseus graph (including hits via the mirror sites) shows accesses to the Perseus main Web server by month. It does not count hits from logos and navigation icons, nor accesses within Perseus itself. The LC graph covers hits from American Memory and THOMAS as the two were targets of this research. LC statistics record hits from the image files, the HTML or SGML pages, and the results of the CGI program, but not valid indications of the number of visitors or number of pages viewed at a Website. Michigan provides users with statistical reports based on each collection, arranged by common access categories, that is search, browse and view categories. For example, in the search category, it includes simple, Boolean, proxy and other searches; in the view category, it includes hits from TOC, text, page/image, thumbnail and so forth. The problem is that Michigan statistics do not provide the total sum of the results for all its collection but collection statistics broken down by month. Since there are more than one hundred and fifty collections in Michigan, the Michigan graph was then estimated based on monthly data from the top fifteen collections with the highest hit rates during the year 2000-2002.

Note the low points appearing in the Perseus graph, which goes up and down at the beginning

and end of the school terms, indicating heavy usage by the education community. The graph shows that Perseus has become an important resource in the Humanities area, especially the Greek and Roman materials which, according to the Perseus Website, form the most requested collections. The Perseus graph also indicates the practical values of the materials Perseus has. Taking into account the fact that Perseus has the least staff and resources, I suggest that Perseus is a success when compared to the other two case studies in terms of usage statistics.

Michigan has the lowest usage statistics among the three digital libraries. This may be because the material is intended primarily for users in the institution; also, the material is highly academic and not oriented towards the general public. Other possible causes of low usage might be strict access management and a preservation-oriented mission rather than the focus on a wider audience than their own campus users (students and researchers) with highly specific needs. Additionally, Michigan does not make dynamic links within the database in the way that Perseus does. So the resulting materials have less added value. From the research interviews, I noticed that there is a certain lack of promotion because the University Library culture is such that marketing is not regarded as important or worthy of effort. My view is that more marketing would help to make Michigan materials better known and thus boost usage.

The LC graph shows lows in summer breaks as well, although they focus on the general public as their main audience. This indicates that the Learning Page project, an on-line gateway specifically tailored to the needs of students and educators, was a successful vision. LC has the largest collection and leads the way in cultural innovation and learning. The size and scope of the LC collection result in the highest number of hits among the three case studies. Most probably, there are users who find the LC Website front page for other reasons and are attracted to these projects by serendipity, since these projects are only one click from the front page. This is another reason for the high usage of the LC collections. Incidentally, there was a large rise in the autumn of 2001 which was maintained until the end of the period under consideration, December 2002. There is no clear explanation for this, though there has been around the world a large increase overall in Internet access, and I think that this large increase in usage in the case of LC may be merely a reflection of that. Alternatively, the increase could be due to publicity from sources such as magazine articles. Another reason for this could be the 9/11 incident which increased the patriotism of the American people.

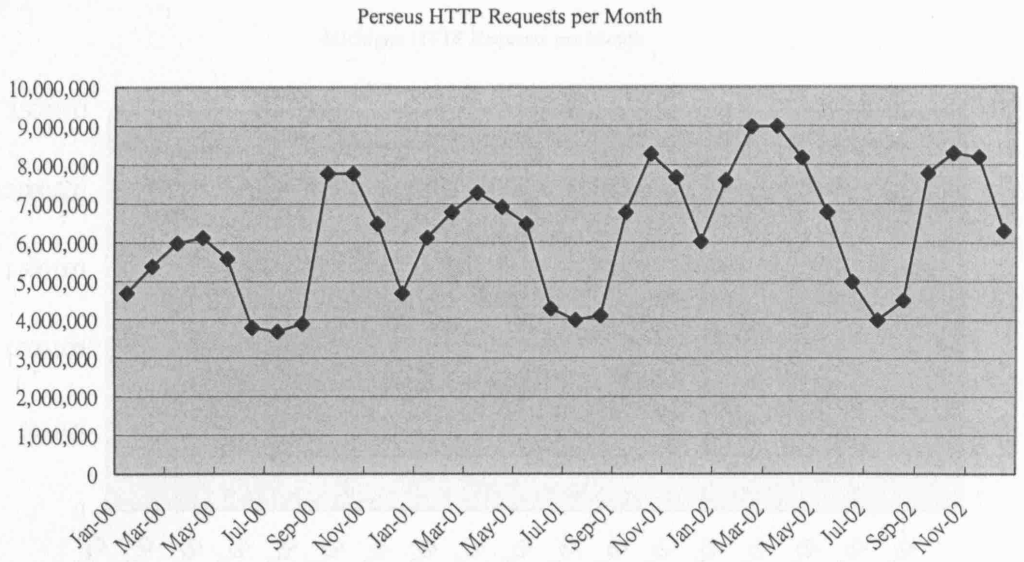

Figure 12: The Perseus HTTP requests per month

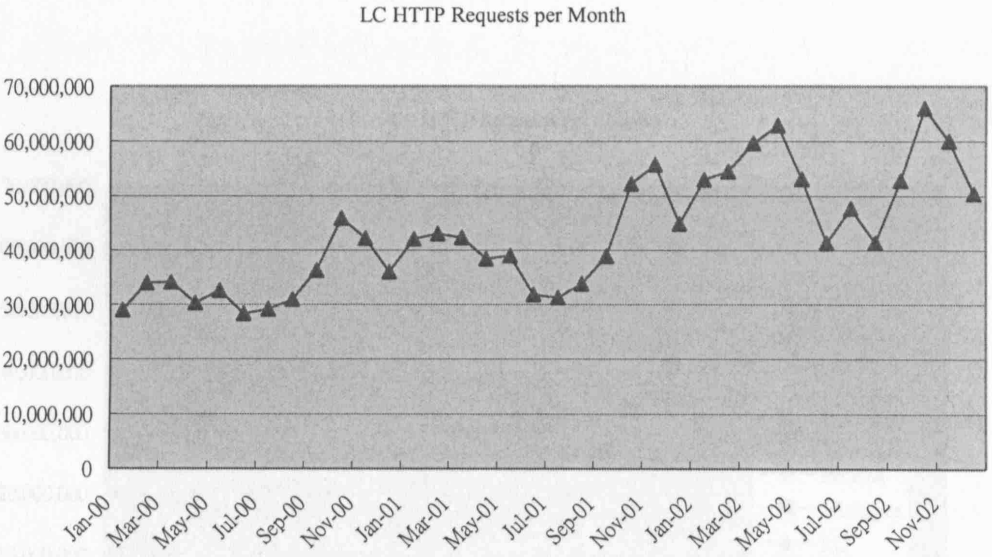

Figure 13: The LC HTTP requests per month

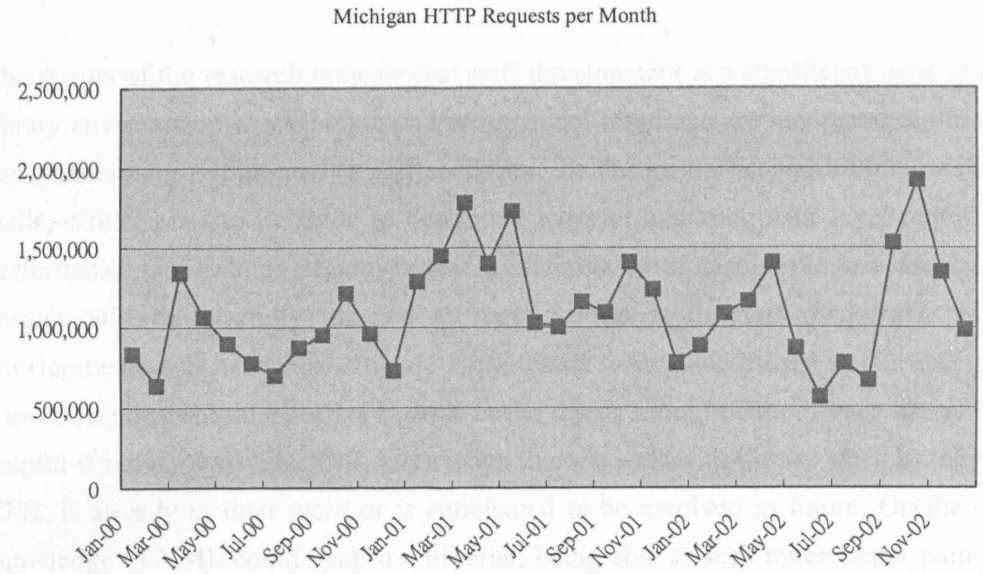

Figure 14: The Michigan HTTP requests per month

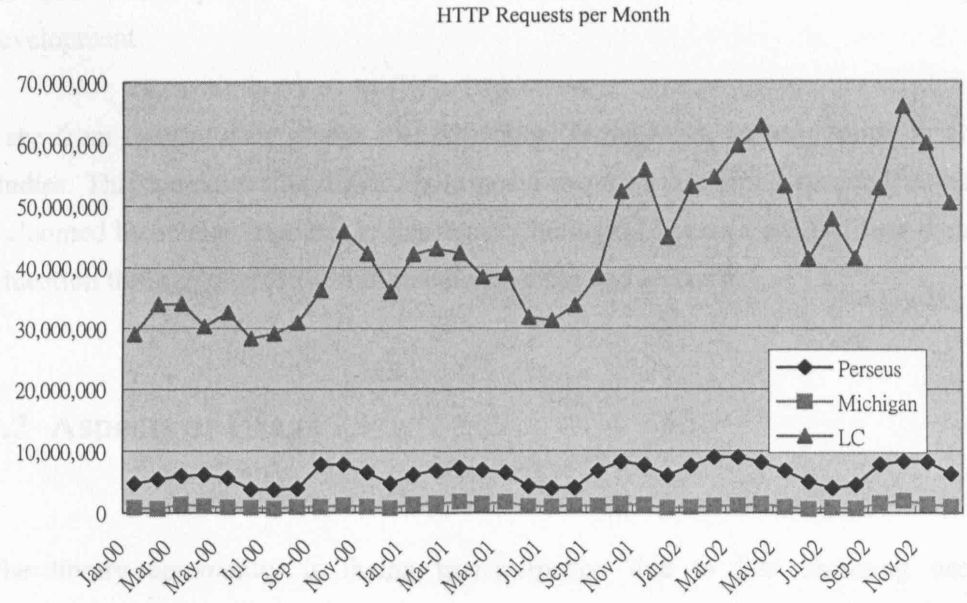

Figure 15: The three digital libraries HTTP requests per month

9.1.2 Conclusion

The results of the research indicate that staff development is a significant issue in a changing library environment as well-equipped professional librarians are key resources to developing and maintaining a high-quality digital library. To ensure the maximum exploitation of staff skills, efforts need to be made to determine ways of assessing skill level requirements and performance and training effectiveness. A continuation of specialized and dedicated training periods will be carefully planned to meet the needs of staff re-skilling. All of these developments will have their ultimate aims, that is to enhance library staff's ability to deliver a more responsive and effective service to the direct benefit of the library users. It would be helpful if library staff had XML knowledge to cope well with library work as they know that XML is already in their work or is anticipated to be involved in future. On the other hand, knowledge of XML could lead to a librarian being able to seek much better paid jobs in the commercial sector, since XML has many commercial applications. This could apply equally to computing staff.

Open source has great potential for economies, enabling sophisticated applications to be built at a reasonable price allowing a good return on investment. Reliable and feature-rich software and support networks allow the possibility for digital libraries to build up successful Web services. Thus open source is a critical factor in the cost structure of digital library development.

I see from the statistics graphs that the school terms make an impression in all three case studies. This indicates that digital multimedia materials are being adopted by educators and welcomed by school learners. In this sense, the digital libraries are fulfiling their mission of education through interactive multimedia teaching and learning.

9.2 Aspects of Usage

The library community is facing transformation due to fast changing user demands. Increasingly, because of the need to optimize services made possible by digitization, the library community ought to know how users approach the discovery and use of digital collections. The issues covered here include the design of interface systems which lead to the resources being presented meaningfully to users, especially given the changing behaviour and needs of active users. In fact, there is clearly a close relationship between resource creation, resource access and resource use. Thus, the digital library evaluation scheme would be based

on a broad view of the subject area.

Maintaining an ongoing digital library evaluation scheme is important to digital library development. A digital librarian will need to know a great deal about how users approach the discovery and use of digital collections and what that means for how they are presented and the kinds of finding aids that are provided. Managers need tools to help them evaluate the usage of their systems. This may lead them to select new software or a new system, but they need the tools to justify that the level of usage requires additional expenditure. Librarians also need the evaluation results to manage the content which is most needed by users. A digital library will need to generate a user community, so that the need for the digital library continues and the budget is not cut. Questions for investigation include:

- What is the benefit to users of XML?
- What feedback do you have from users?
- Have you developed an evaluation scheme for digital library services and collections?
- What kind of approach to DL evaluation have you developed (for example, evaluating systems, interfaces, or evaluation from the users' point of view - user needs, preferences and user community)? How far have these gone?

In my research, it would have been interesting to have direct contact with users, but in a digital environment, where users may be distributed anywhere that the Internet is available, it is difficult. As stated in the Methodology Section of Chapter 6, in order to derive some data from the users themselves, I had prepared user survey questionnaires for my three case study digital libraries, but these were rejected by Michigan and LC, and then it was decided not to pursue Perseus for a user survey. I therefore had to rely on information from the case studies themselves.

9.2.1 Discussion

9.2.1.1 Users/Usability

1. Interviews

Perseus

The Perseus environment includes texts, still images, QuickTime, the Virtual Reality Modeling Language and computer models, and because of XML this allows users to explore the three-dimensional reconstructions of historical sites.

A member of Perseus research staff stated that the Perseus community, particularly in the field of classics, is lively and demanding. As stated in the Website, there was an average of twenty mails to the Webmaster each day. And in order to provide quick and active feedback to the users, the Perseus Webmaster created a Frequently Asked Questions (FAQ) database and some tools for managing messages including automatic responses. These emails have been recorded and classified. It is Perseus' goal to use the Webmaster database to make improvements not only to the content of Perseus, but also to the "help" documentation.

Michigan

Michigan has a relatively small amount of image collections which use XML. The image collections (XML), together with the large amounts of text collections (SGML), allow users to conduct structured searches throughout the database.

Michigan receives feedback from all different types of users covering different areas of questions, but mostly they are about the system and content. The questions are mainly from the institutional members of DLXS, who tend to ask about system installation; also, there are relatively few questions from users on campus, as there are core groups of users who use online resources for teaching, learning and for research. The user feedback is dealt with by staff who firstly filter the context and forward the questions to appropriate staff members. For example, staff involved in MOA were the selectors who initially selected the materials, and they are responsible for the MOA questions. Michigan keeps tracking the feedback; approximately fifty questions a month, but does not yet have a plan for a FAQ database like Perseus has. According to the research interview, there has been high satisfaction from campus users in supporting teaching and learning on campus because this has been the main purpose of the University of Michigan Digital Library, so that the investments taking place in digitization projects on campus collections could be realized.

The Library of Congress

I was told that the Library of Congress has used XML in OAI, where XML was proving very valuable for sharing metadata among institutions. However, OAI does not benefit users directly, so the interviewee at the Library of Congress thought that XML was involved in situations where XML was useful as a convenient carrier of the metadata in OAI.

I learnt that there was a big usability exercise for the Library Website, but was told that there was no connection between that and digital library services and collections, and it did not cover American Memory. Since I could not get any information on user feedback, and was not pointed to any unit which could possibly provide me this information, on 7 September 2004 I communicated in real time with a Digital Reference Librarian through the "Chat with

Librarian" service on the American Memory Website. I learnt that the American Memory/Digital Reference Team at the Library of Congress receives on average between 1100 and 1300 inquiries of various types per month. This number includes an average of roughly 150 "live chat" inquiries which tend to be research or general interest questions from patrons seeking additional information on materials in the online collections, or in the Library of Congress in general. The user feedback is generally on many topics, such as content of the collection, though occasionally it is on site navigation and searching, possible errors encountered in the text and so on. The feedback is forwarded to the appropriate Division for response. Also, the feedback is tracked and archived to official reports on reference "traffic". There is a reference team who see all the feedback submitted by patrons using the Library's "Ask A Librarian" service.

2. Analysis and conclusion

In Europe, some useful work on user statistics has been done in this area. One example is the Society of COllege, National and University Libraries (SCONUL) Advisory Committee on Performance Indicators. One of their tasks is to help to improve the management of libraries and information services via partnerships with other organizations in other countries. In 2001, they extended their collection of statistics to include those related to the use of digital materials. [SCONUL, 2005]. Another example is the EU-funded EQUINOX project, which addresses the need for all libraries to develop and use methods for measuring performance in the new networked and electronic environment alongside traditional performance measurement, and to operate these methods within a framework of quality management [EQUINOX, 2000].

As a supplement to the information gained in the research interviews, in January 2003, I reviewed the features found in Perseus via a search example as a novice with a limited Latin background. I did a search for a specific word, "thief" in Plato's *Republic*. I found Plato's *Republic* via the Collection Viewer searching on Collection Contents and Texts. When I searched the actual text of Plato's *Republic*, using the look-up tool, I got no hits because the text was the Greek text (even though I had specified English text search). When I searched via the English Index tool in the 'Greek and Roman materials' collection, I retrieved many instances of the use of the word "thief", in fact 342 instances in 102 works, and had to go through these to find Plato's *Republic*, where I found the reference I was searching for. Afterwards, I realized that in the list of texts, Plato's *Republic* appeared twice, Greek text and English text, so when I searched the English text, I found the correct reference immediately.

The result of the search is shown below.

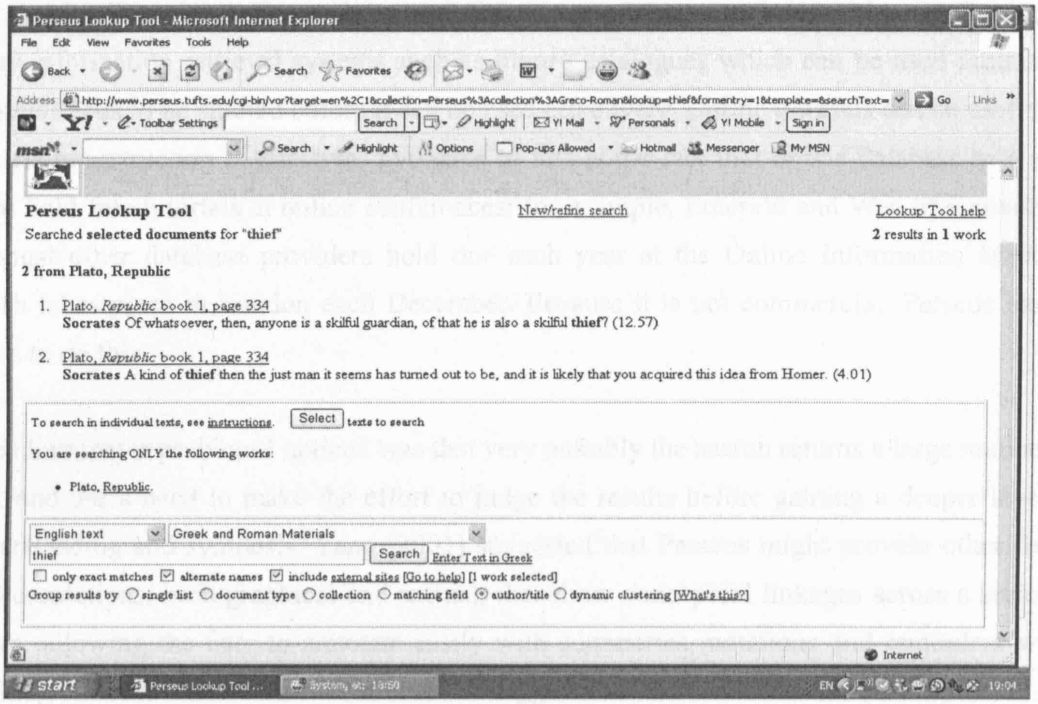

Figure 16: A search for "thief" in Plato's *Republic*

The "help" function gives the example of how to find images with dogs. I did the same search for "cat" using the Art and Archaeology Index for keywords in the category of object: vases. There is a keyword 'animals' under which 'cat' is listed and I was led to an image of a vase portraying a cat-like animal in the associated text of which was the expression 'cat or panther'. Given the difficulties of searching through images, because this type of search always depends on a textual description which may not include the word for which you are searching even though the concept represented by that word is present in the image, this seemed satisfactory and the search was successful.

Virtual reality tools require high bandwidth and Quicktime on the computer, and the search through 'London, King William Street' retrieved still images, but I was not able to see the panorama or the movie. This is an indication of how one of the features of Perseus is not readily available to educational institutions such as schools for which this tool was intended, since many schools have low to moderate bandwidth.

One of the major problems I found which was caused by the Perseus dynamic searching environment was that often I got lost through linking and did unnecessary printing ~~because I was not sure if I could find more interesting materials by further linking~~ when deciding the

usefulness of the materials. This would be less of a problem if Perseus visualized the search results and provided the layer links feature, so that users could always come back to the top layer and follow the different links in different collections. Another problem is that unlike many information retrieval systems such as library catalogues which can be used intuitively, Perseus needs to be studied before it can be used effectively. Online tutorials can be useful but classroom instruction is effective. Evidence of this is the fact that online database providers often hold free tutorials at online conferences: for example, Emerald and Web of Knowledge amongst other database providers hold one each year at the Online Information Meeting, which takes place in London each December. Because it is not commercial, Perseus has no funds to do this.

One more major problem I noticed was that very possibly the search returns a large number of hits; and users need to make the effort to judge the results before gaining a deeper level of understanding and synthesis. Yang [2001] suggested that Perseus might provide other forms of presentations, both graphical and textual, that show conceptual linkages across a series of nodes, allowing the user to annotate easily with summaries, notations and reminders while working with the links. This is similar to the suggestions that I made in this section.

In Perseus, the links are provided automatically for every document when it is displayed. Because subject terms are linked to dynamic hypertexts, not to simple glossary or dictionary entries, readers can explore types of material they might not have thought relevant. Another benefit is that new documents appear on pages automatically once they are encoded in the system database. The example of such a feature is "citation reversal", which happens automatically each time a text is added to the system, which is one of the distinguishing features found in Perseus.

Michigan thought that fast loading of pages was essential to reach the widest possible audience, so they used fewer icons and graphics than other Websites. Michigan separates the searching interface into text and image collections, which makes for efficient search results, but separating them in this way is rather inconvenient from a user's point of view because they have different search interfaces. Both of them are organized by group collections.

UMDL Texts is the central access point to the electronic books and journals provided by the Digital Library Production Service. As of January 2003, there were, according to their latest updated Website, a total of forty-nine collections serving more than thirty thousand Texts [UMDL, 2003]. Some collections are restricted to use by authorized users only, such as university staff, students and so forth. In the UMDL Texts search menu, users can view the full descriptions of the collections provided in the Groups Menu, which is informative. Most

searches in UMDL Texts allow users to select from among one or more "regions" (that is, author, title, citation) or distinct sections (that is, lines, pages, chapters, paragraphs and so forth) of a work. UMDL Texts support a simple search with which users are able to search for words or phrases anywhere in the Texts. Additional features include truncation and three advanced searches. A bookbag option, which I thought useful and user-centred, allows users to select, collect and email citations, enhancing the search interface which could be criticized as being plain and traditional. Michigan has plans to provide more search types in the future.

I notice that the Michigan DLXS user community is active. Thus, it can be seen that one of the DLXS institution members, Tobacco Control Archives (TCA) at the University of California San Francisco (UCSF), was impressed with the performance of the DLXS search engine, which met most of the project's requirements and allowed the project team to meet its objective of building the library in one year [Schmidt et al., 2002]. The UCSF team recognized that the extensive collegial support and feedback from the Michigan DLXS was beneficial and smoothed the implementation of DLXS technology.

The LC American Memory includes features that permit users to search across collections or an individual collection, as the Library recognized precise retrieval is more important than efficiency [Arms, C.R., 1999]. LC also introduced tools to help readers find materials efficiently. Browse Collections is designed to give users a useful clue in finding collections of interest to them by name of collection, topic, original format of materials, time period, by particular region of the country or international. The users can search on full text (in some instances) or bibliographic descriptions, where available.

In the past, LC had as its core mission supporting the Congress to gain quickly relevant and verifiable information, but now LC has an institutional challenge to broaden their practices and have made resources available online to the general public, especially to young students in schools with slow network connections. LC Learning Page offers organized help for searching the Library's primary resource of rare Americana collections, which are categorized by the Events, Topics, People, Time and Places of American history. Teachers take advantage of Features and Activities and Lessons Plans sections in planning their class materials using digitized films, photographs and sound recordings which engage students in new way.

Making the collections more accessible should be the priority policy for all digital libraries. Therefore, a friendly and helpful search screen is essential. I found that LC and Michigan provide a rather traditional searching environment with limited tools, whereas Perseus, using XML linking technology with a collection of tools, provides a full content search, drawing

results from the entirety of the varied types of source content in its database. However, the novice user might not find Perseus so easy for developing a useful set of strategies for searching information in the database because of the complexity of the tools.

The three case study digital libraries take into account user feedback. Michigan particularly takes feedback seriously, as the DLXS User Group is one of their main sources of revenue.

9.2.1.2 Evaluation Schemes

1. Interviews and Supplementary Research

Perseus

I was informed that during the early stage of the development of Perseus, Gary Marchionini at the University of North Carolina at Chapel Hill conducted an ongoing evaluation component as a research activity in aspects including technical issues such as searching, browsing, collaboration, authoring and metadata; social research such as user needs, sharing, community development, formation quality of training and learning/teaching in digital libraries. Marchionini's team provided a useful external perspective on the work of Perseus within the project that was central to its progress [Marchionini, 2000]. At the time of interview, Perseus continues to seek cooperation in the subject of user studies with the University's Psychology Department.

Michigan

During the research interview, the interviewee said that they had not set up a general evaluation scheme to evaluate the full range of operations as viewed by their users. The evaluations which have been completed are mostly project-based, such as project MOA in user needs, or assessment on functionality and usability on specific projects. Michigan expected to have a standardized and formal evaluation programme as soon as they recruited the right staff. In addition, Michigan has been seeking the possibility of cooperation with the School of Information at Michigan as a research project, but unfortunately they have not found research topics of mutual interest.

The Library of Congress

LC does not have an overall methodology for evaluating digital library services and collections, but rather evaluates projects individually. I was told that it was a problem to evaluate services that they provide to satisfy the needs of users that they never meet. Nevertheless, LC participated in the DLF activity on digital library use, users and user support programmes in which a number of assessments and surveys have been conducted among DLF members.

During the period when LC NDLP was proceeding, a team from the Human Computer Interaction Laboratory (HCIL) at the University of Maryland had been working with a team at the Library of Congress to develop and test interface designs for NDLP. This collaborative effort aimed to create user-centred interface prototypes [Marchionini et al., 1998]. A number of design goals were outlined according to principles, proposing that users should maximize their interactions with information resources and minimize their attention to the system itself, and that both browsing and search strategies should be supported. Additionally, three types of tools were developed especially for overview, preview and gathering collections.

2. Analysis and conclusion

Digital libraries support their use in ways that reflect the new information environment, and the changing behaviour and needs of active users. For Human Computer Interaction (HCI) specialists, areas for investigation include user behaviour, service contexts, user interface, reuse and exploitation. HCIL at the University of Maryland highlighted four research and development issues related to digital libraries in general as the result of its work, from which I found a substantial influence on the later development of digital library and digital library evaluation schemes. HCIL suggested digital library staff must carry out regular user needs assessment, ongoing usability testing and iterative design procedures in order to develop and test principles and guidelines for user-centred digital libraries, as digital libraries will grow and change with users and technology. Secondly, digital libraries must serve diverse user communities and need to develop appropriate interfaces for varied users and needs. Thirdly, the development of the various tools and the highly interactive environment bring issues in the toolkits needed for digital librarians. Finally, technology is not exclusive to the development of the digital library. Digital libraries should prepare for challenges from social and political factors, as digital libraries are rooted in organizations which are themselves also affected by new trends in information technology and a greater use of the Internet.

The evaluation and investigation of library technology is not a new research area [TREC, 2003]. .. Human Computer Interaction has different usability evaluation methods. HCI may be used in order to evaluate the usability and levels and types of usage of digital libraries and kinds of users. Fuhr et al. [2001] research group claimed that digital library evaluation deals with the effects of digital libraries on subsequent human information behaviour; therefore, the evaluation should be process-oriented and iterative rather than product-oriented and summative. Blandford [2004] thought digital libraries need a "toolbox" of evaluation techniques that could possibly address different aspects of evaluation challenge, and thus could be useful for digital libraries at different stages of the digital library development process. These theories match the research findings of Marchionini as I discussed above.

Perseus and LC saw user studies critical to their future development. They had been focusing more on user-centred evaluation schemes than on technical systems. Perseus invited an external evaluation research group to conduct an evaluation scheme from early on in the project. The heart of the evaluation scheme was instructional technology [Marchionini, 2000]. This reflected on the classes of evaluation objects: learners, teachers, the technical system and the content. At LC, a large-scale end-user evaluation was done before the American Memory Pilot was launched. LC considered serving the nation's schools as potential audiences, and worked with teachers and students to plan materials they would find most useful in electronic format. Michigan has done relatively little on user evaluation but focuses more on DLXS user group activities where they organize workshops and meetings, and make active contact among member institutions. This is because the DLXS user group generates ongoing revenue for Michigan.

When comparing the three case studies, Perseus has done relatively more on user evaluation than the others. I thought this could be because Perseus depends heavily on a grant, so they need to put their effort into every possible aspect of the digital library. Michigan did the least evaluation of the three, and had the lowest usage statistics as discussed in Section 1.1.3 of this Chapter. It could be because they were afraid to receive a negative evaluation which would not help them to justify their existence. Indeed, when requested to support the author's user survey, they were the first to reply in the negative.

In short, this evaluation of the digital library will be ongoing and based on a general view of every subject area. I found my three case studies had not put enough effort into this area.

9.2.2 Conclusion

The evaluation of digital libraries is essential for future development. However, the importance of the evaluation of digital libraries has not been fully recognized. Digital libraries increasingly embed resources in wider interpretive contexts. It could be curricular materials, guiding materials or academic materials. These may be realized through structured, searchable, sharable documents that provide instructional, learning, navigational or other interpretive narratives for services; therefore, digital libraries will need to select a well organized evaluation scheme in order to assess user information needs and corresponding tasks such as evaluation in the area of teaching and learning, in particular the evaluation of the real benefit to the users of implementing XML in the system.

9.3 Organizational Aspects

For most organizations which have digital libraries, it is probably the case that they are not capable of setting up and managing sustainable digital collections without assistance from outside the department which hosts the library. This issue is particularly obvious when creating collections with large numbers of documents rather than occasional documents. In the higher education sector, the challenges of the digital library do not relate only to the library. They belong to its host institution and need to be resolved at an appropriate institutional level.

Libraries are in danger of not being able to afford to make their collections readily available in the information and knowledge society. So, when building up a stable supportive infrastructure, digital libraries may consider establishing a number of relationships and partnerships within the subject communities and the information science professions.

At the heart of the digital library programme is the goal of providing broad networked access to resources using innovative technology and taking advantage of collaborative work on standards development across participating communities. It can, in fact, be a good opportunity for exploring new technologies as digital libraries will encounter a certain amount of technological challenges that will lead to significant innovation.

Implementing XML in digital libraries is an innovation as the technology is new. XML technology could be implemented in areas such as content creation (XML based metadata), content management (XML native database) or even an open infrastructure under a common language like XML, which facilitates delivering information smoothly on a variety of platforms. Questions under these perspectives include:

Institutional strategy

- What partnerships and collaborations do you have? What is the rationale for these?
- What are your models for sustainability? Where did the original funds come from?

Future work

- What are the major contributions of your Digital Library?
- What are the future directions?

9.3.1 Discussion

9.3.1.1 Partnership and Collaboration

1. Interviews and Supplementary Research

Perseus

Perseus has active connections with a number of Humanities institutions based in Europe and in the United States which enrich its collections with broader subjects in the Humanities. For English language collections, Perseus worked with the Modern Language Association. For the London Collection and the History of Tufts University Collection, Perseus worked with Tufts University Archives. For the Boyle Papers, Perseus worked with the Max Planck Institute for the History of Science in Berlin. Perseus included third party materials to test the applicability of its tools to a broader range of materials and to provide access to particular services within the Perseus Digital Library, for example the Library of Congress American Memory collections on California and the Upper Midwest. In addition, Perseus made available its XML document manager to the Stoa Consortium at the University of Kentucky, which is a centre for electronic publication of classical scholarship and, in return, Stoa publishes editions of Perseus' texts. Furthermore, Perseus is using the Open Archives Initiative protocol to share XML based metadata and software tools with members in the Open Language Archives Community (OLAC).

Michigan

Michigan has active internal and external partnerships and collaborations. In the University, Michigan has partnerships with the School of Information on faculty research as a number of Michigan's collections have been included in the testbed of the NSF/ARPA/NASA-sponsored Digital Library Project of the School of Information. The Program for Research on the Information Economy (PRIE) is another instance of cross-unit collaboration between DLS and the University departments. In addition, the University of Michigan Press has served as a licensing or billing agent for DLS digital content and services provided to other institutions. For example, the Press may publish or license a hyperbibliography of Middle English text resources to other institutions. Michigan has long and rich partnerships with publishers. For example, the Elsevier Science TULIP project delivers materials in electronic journals in the field of science; an SGML version of Grolier's Encyclopedia of Science and Technology was created and deployed based on a cooperative project; the electronic version of the Human Relations Area Files is another cooperative effort.

Michigan also collaborates with a number of external institutions under a number of grants. For example, being a partner with Cornell University through the MOA project; being a host

for the Library of Congress for the Abraham Lincoln Association. Meanwhile, Michigan provided elements of MOA models to the LC Preservation Directorate for digitizing a ten-volume journal, *Garden and Forest*. This journal is now online through a collaborative effort between the two institutions. In addition, I learnt from research interviews that there has been an encouraging collaborative effort based on OAI technology. Michigan has been experimenting with the University of Illinois at Urbana-Champaign (UIUC) and other institutions on two ways to transform data through OAI and support their DLXS software, making partners' repositories available online into Michigan's search interface.

The Library of Congress

The Library of Congress has been participating in many partnerships and collaborations. The rationale for this is to promote standards and provide wider and more effective access for the educational community to the Library's historical collections.

Firstly, collaboration to build collections of selected digitized historical materials hosted at LC: the LC/Ameritech competition involved agreements on 23 awards to institutions. Similar arrangements have been made with four or five other libraries and historical societies in the United States; collaboration with foreign libraries to build collections with subjects related to exploration, immigration, and other interactions with what is now in the United States.

Secondly, collaboration for contributing historical materials digitized by LC to other virtual collections: RLG's Cultural Materials resource; RLG's archive of finding aids; other libraries using the Open Archives Initiative Protocol for Metadata Harvesting to harvest catalogue records for items in American Memory.

Thirdly, cooperative cataloguing collaboration: LC organizes partnership programmes for cataloguing monographs and serials with records following established guidelines (more specific than AACR2) and contributes to OCLC and other bibliographic utilities.

Fourthly, being a member of organizations to promote standards and community practices: W3C (World Wide Web Consortium); DLF (Digital Library Federation); NISO (National Information Standards Organization, Z39); MARBI (US MARC standardization body).

2. Analysis and conclusion

Collaboration through partnerships has emerged as an important strategy for libraries in creating value added services and enriched interactive environments [British Library, 2003]. The collaborative model could be internal or external; it could be between academic

institutions, research institutions, foundations, government agencies and industrial partnerships. One area in which libraries cooperate is in training. This is largely because all libraries ultimately serve the same functions and they tend not to compete with each other even if their parent organizations are competitors. In the areas of new technology and training for new technology, they are quite proactive in setting up cooperative training courses. Training in XML commercially can be very expensive, but libraries can usually organize training at lower cost than the commercial world or they can share in training through professional associations such as CILIP.

Mutual benefit is the essential reason for the partnership. A partnership can achieve things which its partners cannot do alone. This can lead to innovation as the expertise is combined; also, this can lead to financial savings based on efficiencies. The benefits to each partner may be articulated explicitly. The investment in software and hardware can be a practical reason to seek collaboration. Collaboration has its potential to share expertise as the three digital libraries have identified earlier, encouraging innovative thinking and thus improving problem-solving. Another essential advantage in partnership is that resources can be better attracted, allowing broader resource inputs and thus provide better services. The rapid evolution of new information technologies accelerates the concept of collaboration in digital libraries. The OAI is a good example.

Partnership is even more crucial in the library community as a whole. Libraries enter into partnerships mostly within their own sector, for example libraries cooperating on union catalogues, Library of Congress Cataloguing-in-Publication Data, OCLC online shared cataloguing system and M25 Consortium of Higher Education Libraries in the United Kingdom [M25 Consortium, n.d.]. Experts estimate that fewer than ten percent of all cultural heritage institutions in Europe are ready to participate in the digital era [European Commission, 2002]. A great majority of memory institutions do not possess the human, financial and technological resources to participate in the Information Society. Archives, libraries and museums could seek strategic partnerships with industry or institutions in order to attract resources, minimize risks and increase expertise.

In Europe, the European Commission funded The European Library (TEL) project, which was set up to develop and test open standards, working methods and practices that can readily be adopted by all national libraries working as a seamless partnership [TEL, 2003]. It was hoped that this co-operative framework would lead to a sustainable pan-European digital library based on distributed major national and deposit digital collections held in the participating libraries and agencies.

I noted that because Perseus has limited resources and financial support; therefore, they had a relatively strong marketing ability and an aggressive partnership and collaboration worldwide that were not found in the other two case studies. Perseus perceived that the combination of intellectual opportunity and collegiality is extremely important for them.

In an increasingly complex digital library world, much progress is built upon powerful partnership and collaboration, a long-term relationship, a sense of trust and respect and the free exchange of ideas within groups of individuals or across different disciplines and institutions. It is fair to say that successful partnership and collaboration contribute to the successful digital library initiatives of the 21st century. The three digital libraries are good exemplars of such.

9.3.1.2 Sustainability

1. Interviews

Perseus

Perseus grew up with grants, as they have limited support from their mother organization. I was told that the creator of Perseus has constantly been worried about finance. Perhaps because of this, Perseus has been vigorously building up partnerships with other institutions and doing marketing to compete for external funding. Ideally, Perseus views these collections as the equivalent of common cultural heritage which is traditionally supported by states; people are free to access these precious materials. In this sense, it is difficult for Perseus to identify models for sustainability that support the original mission as a nonprofit entity. Perseus libraries are designed to maintain both core and in-depth information over a long period of time, so the solution would possibly be to incorporate the work they have done into the University Library as a long-term model of sustainability.

Michigan

Michigan foresaw the importance of sustainability and partnership management in digital libraries. As discussed in Section 3.1 of Chapter 6, the TULIP, PEAK and JSTOR projects explored the electronic commerce systems in a digital library environment. The DLXS was designed to cover ongoing development, maintenance, and support activities directly associated with the DLXS through its licensing operation. The main revenue comes from the profits of the "production factory"; that is, DLPS provides digitization services (primarily in the Humanities), not only to support a variety of projects from the University Library but also from the library community, where there is a significant demand in transforming original materials to digital format.

The Library of Congress

LC has the privilege to have the least financial worries among the three digital libraries. LC fortunately encountered fewer challenges on funding when compared with the high expectation to speed up digitization. The initial funding for digitizing historical materials at LC came from a combination of special congressional appropriation and funds raised from private-sector donations. The private enterprises like to have their company names appearing on the LC digital library Website.

The initial funding for the LC DAVPPP was from a private donation to put the building in place at Virginia. The matching funding from the Congress will contribute to the prototyping, systems development and digitization. Overall, private funds form the largest contribution to sustainability.

2. Analysis and conclusion

Digital library development demonstrates its potential in terms of content creation and collaborative partnership. However, it will take time to achieve a critical mass, and that sustainability is critical for fulfilment of its potential. The possible challenges include sustaining the confidence of partners (or members) and external funding agencies in seeking more funds and support, and in ensuring a steady supply of staff into the future.

Not until 2001 has the importance of sustainability been addressed by numerous discussions and surveys done by research organizations [CLIR, 2001; National Science Digital Library, 2003; Zorich, 2003]. Yet, these discussions pointed out that more practical sustainability models were needed to be demonstrated as examples of good practice to be followed, and that other issues of sustainability needed to be explored.

Most digital libraries struggle with acquiring funds to retain key staff and at the same time find a sustainability model difficult, particularly in identifying a source of ongoing revenue. Doing work successfully while the initial funding is available is challenged by creating a sustainable model for after the grant runs out. A number of local factors may influence fundraising including the nature of the library, library environment and institutional culture.

I suggest that Michigan demonstrates a successful model in collaboration with research partners on research projects and consolidates research results as the stages of the digital library development progress. Moreover, Michigan integrated and transformed the research results to upgrade its digitizing services to the customers within the library community outside the University, and by this means they are able to have a reliable funding source which is necessary because they do not receive sufficient support from the University Library

to maintain sustainability. Perseus is the exception among the three case studies. The creator of Perseus takes the sole responsibility for sustainability which involves attempting to tap every possible funding source in the United States and Europe. For a digital library project of this kind, there could be a potential risk of failure if grants were not continued.

Building a digital library can be a costly and lengthy process. Funding is an ongoing issue for many digital library projects. It is fair to say that support from public organizations and bodies and private sponsors and contributors is instrumental to the success of digital library development. The Library of Congress NDLP is the best case.

9.3.1.3 Contributions and Future Directions

1. Interviews

Perseus

The view of Perseus staff was that the new generation of students learning Greek and Latin have grown up with Perseus, which can be regarded as one of its major contributions. At the same time, the nature of Perseus' work has revolutionized how scholars in the Humanities think in terms of what implications technology had for them. Perseus is making good contributions to the transformation.

The general structure is in place. Perseus wishes to keep pace with on-going research issues through partnership and collaboration. Future work will focus on new enquiries in new kinds of information, linking other sources of data, which might be found useful in doing research, such as conventional indexes. There will be more technologies on linking quotes and citations, tabular information, monetary sums, temporal spatial querying and providing link services to external datasets. XML will remain at the core of document handling system.

According to my update emails with Perseus on 8 September 2004 [Mimno, 2004], I learnt that they are preparing a new version of their document handling system; more recent XML initiatives are expected to be implemented in the infrastructure. Firstly, Perseus is now using XSL both for texts and for X3D models. Secondly, Perseus will be creating full METS records in order to transfer their digital objects to Tufts Digital Library (run by the Tufts University Digital Collections and Archives), which is based on the Cornell FEDORA repository (METS based). Thirdly, Perseus expects to implement XML Web Services as the base of a broader array of new services. A Web service for morphological information, based on XML Web Services technology, has already been mounted through a partnership and collaboration project, Cultural Heritage Language Technologies (CHLT) [CHLT, n.d.]. Other XML-based services are in the pipeline. One of the new services will enable the user to find

texts on the Perseus database by entering a sentence like this: "Get me Caesar's Gallic Wars, book 2, chapter 10.".

Michigan

Michigan thought its major contributions were sharing experiences and resources with other institutions partly through DLXS, through hosting of collections for institutions which were not able to host their own for whatever reasons such as lack of facilities, expertise and so forth, and through partnership.

Michigan had been working with standards and recommendations such as SGML, XML, and TEI and now more XML was anticipated. Firstly, the new library management system, Ex Libris, had XML capability. Michigan expected that the system would do things differently in the future, when they would be using XML in different areas of the library. According to my update emails on 17 August 2004, Ex Libris had been up for less than a month, and they were still only just beginning to integrate the digital library processes into their workflow [Powell, 2004]. Ex Libris solves the special character problems as expected. As far as DLXS was concerned, Michigan continued to improve DLXS middleware to meet the needs of member institutions. Michigan planned to continue to strengthen the XPAT as a low-cost, high functionality especially regarding XML and SGML awareness. Michigan confirmed via email that XPAT was able to index UTF-8 (an encoding that provides access to the Unicode character set). Michigan OAIster has this functionality and is completely XML.

Through the follow-up emails, Michigan confirmed a complete commitment to XML and XSLT in the new version of Bib Class, release 12 [Powell, 2004]. All the text creation will be transferring from SGML to XML. All the finding aids that Michigan have online are encoded in EAD 2002 XML. Additionally, Michigan stated that they were in the process of hiring a new programmer to work on METS integration.

Also, I learnt during the research interviews that there will be more future work in collections and services. Michigan will shift the hand-selected projects to more partnership digitization-based preservation. The Preservation Division and DLS Digitization Services have established a collaborative partnership to share expertise and resources. This will lead to several challenges in dealing with different languages in OCR. Also, there will be more digital conversions for external requests, such as from University units or external donors. For example, DLPS is working on technical reports for the Media Union and Business School. Furthermore, interface support for audio and video resources is under development. DLPS are participating in a multi-institutional grant proposal to the Andrew Mellon Foundation to build a distributed digital archive of ethnomusicology materials, including video and audio

resources. Also, one of the DLPS programmes is to assess non-GIS datasets such as statistics data, but this will raise the migration issue and support issue, since currently there are no standard formats in place. With the success of digitization, DLPS wishes to extend its partnership to support more international collaboration. Several opportunities for world cooperation and resource sharing are under discussion.

The Library of Congress

The LC NDLP effort has made available more than six million heterogeneous digital items from more than one hundred historical collections for public access without concern for institutional or national boundaries. These, in turn, are creating the heritage of the future. This is what LC thought was its main historic contribution.

LC has to decide how to apply XML to its particular problems, although there is no doubt that there will be much more use of XML in the Library. The focus will be on where XML can improve what LC has to do, rather than on how LC uses XML. LC sees the future direction as considerably more use of XML in particular areas: areas in relation to gathering together all the metadata for digital preservation; XML-based ONIX metadata for publishing can be used as the record in mainstream workflow and the digital registry system may have metadata in XML. A number of experiments are going on. For example, the Law Library's primary digital initiative, the Global Legal Information Network (GLIN), is exploring the delivery of an annotated XML tagging scheme for the DTD and a migration plan to XML for GLIN database components [Library of Congress, 2002].

From the follow-up emails on 19 August 2004, I learnt that SGML is still the core in LC but, for completely new projects (rather than new collections being added to old projects), XML has been used where SGML had been used in the past [Arms, C.R., 2004]. XML will be one technology component of most new projects either for descriptive records, for structured markup of text, or for representation of complex objects. For instance, the 'I Hear America Singing' (IHAS) project is based entirely on XML, with MODS for bibliographic description, METS for digital objects and XSLT for display. Further XML implementations include the fact that all finding aids marked up in the EAD DTD are being converted from SGML to XML; XSLT is being included intensively in new projects, as are an OAI data provider service and XML-based finding aids; more migration to MODS for some American Memory collections is being considered.

Although NDLP officially came to an end, digitization work is still proceeding through the main part of the Library Services. Digital conversion has become a part of the continuation work in the Geography and Map Division and the Prints and Photographs Division. At this

stage, there is more internal than external pressure in digitization, as divisions in LC wish their holdings to be digitized at the earliest possible opportunity.

I learnt from the update emails that LC and Michigan are not using RDF. It illustrates our conclusion in Section 3.1 of Chapter 5 that RDF is difficult, and that there is relatively low usage in the digital library community.

2. Analysis and conclusion

According to my updated emails with the three case study digital libraries, I see that Perseus amongst the three digital libraries again firstly adopts cutting-edge XML initiatives such as XML Web Services since its nature is a technology-oriented digital library project. As I identified in Section 3 of this Chapter, a digital library needs to affiliate to a larger organization, so that it can share the resources and have more chance of long-term sustainability. Perseus has recognized this and has been transferring their digital resources to the Tufts University Digital Collections and Archives as a permanent home but maintains an independent "view" of their resources through the Perseus Website. However, I learnt that its position of no support with respect to the University Library has not changed.

It is worth noting that it is not because of the availability of XML technology but because of requests from the user group that Michigan plans to support more XML technology. I discussed in Section 2.1.1.4 of Chapter 8 that DLXS member institutions requested XPAT support in METS and XSLT for results filtering. My view is that Michigan is transferring their infrastructure from SGML to XML partly under pressure from DLXS members. Michigan is a good indication that migration from SGML to XML is the direction for the future.

It is reasonable that LC experiments on XML with new projects but at the same time retains SGML as the core in the foreseeable future, as LC holds massive amounts of materials. In the case of Michigan, they also experimented with XML starting with small collections of images, XSLT and they have committed to transferring all text creation from SGML to XML. This supports my conclusion in Section 5 of Chapter 2 that for old digital library projects, changing their retrospective data from SGML to XML is not a surmountable problem because XML is a subset of SGML. However, it would need every aspect of its working operation to be evaluated in depth before they made any changes. The best way, as we see in the case of Michigan and LC, is that they experiment in new technologies with projects or small collections in a small part of their infrastructure rather than transferring the whole infrastructure at one time.

9.3.2 Conclusion

Research organizations such as the Digital Library Federation, Coalition for Networked Information, Arts and Humanities Data Service and the Network Development and MARC Standards Office at the Library of Congress contribute to partnership and collaboration by organizing training sessions, informing one another about potentially valuable new technologies, sharing the results of any local experimentation or assessment of such technologies and distributing standards within the library community. The aim of the partnership and collaboration is to provide the digital library community with best practices with an economic solution.

Givens [2004] would agree with my perception that digital library development in the United States seems to be more active, devoted and enthusiastic than in the United Kingdom in terms of scale and budget, thus the results are more encouraging and fruitful. In the United States, digital library grants and support come from a wide array of organizations such as from federal agencies, academic or charity-based organizations, or from commercial institutions. Compared to the United States, it seems that digital library development in the United Kingdom should have more visible commercial and industrial involvement along with government and academic institutions. At the same time, it would be beneficial if there were more partnerships and collaboration between Europe and the United States in order to share best practices, especially in exploring XML technology in digital library development, on the lines of TEI, which is such a partnership.

The three digital libraries are still a "work-in-progress". LC and Michigan have plans to digitize many of their holdings while Perseus will continue to work as a research project and look for more partnerships to extend its holdings through collaborative efforts. The current work and emerging technologies have marked the continuation of a long-standing vocation and a prelude to some exciting work ahead. They contribute to a research and development programme that has ensured that digital collections have become fully embedded into research and learning in the electronic environment.

Digital libraries have gradually recognized the value of XML and are seriously considering moving to it as one of their important strategic directions in order to follow that technology trend. The impact of XML in digital libraries will grow with time as the technology matures.

Chapter 10

Conclusion

10.1 Preamble

In 1931, the librarian regarded as the father of library science in India, S.R. Ranganathan, published his Five Laws of Library Science: books are for use; for every reader his book; for every book its reader; save the time of the reader; a library is a growing organism [Ranganathan, 1931]. The five laws are generally regarded as giving a clear description of what a library is, and should be doing. Almost seventy years later, Gorman's Five New Laws of Librarianship reinterpreted Ranganathan's truths in the context of the library of the digital age: libraries serve humanity; respect all forms by which knowledge is communicated; use technology intelligently to enhance service; protect free access to knowledge; honour the past and create the future [Gorman, 1995].

Technology has changed the ways in which people communicate with each other. In academic disciplines, computers and networks have become a vitally important medium for the exchange of scholarly information. The most important reason for digital libraries to exist is that they have the potential to extend the five laws of librarianship: information can be shared; information is always available and can easily be made current more readily than by updating the printed word; new forms of information become possible; information can be reorganized; and information can be preserved for future generations.

In the Summit of 1999 ACM Conference on Digital Libraries, the message was already: "Digital libraries will likely figure amongst the most important and influential institutions of the 21st Century". Not only will future digital libraries dramatically improve access to the world's knowledge, but they will also act as "collaboratories" out of which new knowledge is crafted and refined by widely-distributed teams and organizations, knowledge that right from conception is fully interconnected with previous work [ACM DL, 1999].

There is a wealth and diversity of innovation in digital library development. Markup is the core tool to build a digital library because it allows data structures to be exchanged for digital library

items and adequate descriptive and structural data to be exchanged. Markup has also proved to be a flexible tool enabling interoperability and easy implementation of the metadata required for building a digital library. Without markup a digital library is in danger of obsolescence and its contents may not be available to posterity.

Digital library projects first used SGML to build digital documents in a generic format that could be stored in a central repository and be used by more than one application; its structure remained intact whichever program was being used to interpret it. SGML captures structural components which have many advantages, such as more precise retrieval results. Institutions with SGML archives can convert them to its subset, XML. XML technology has been regarded as a key technology in Web applications, as it is a W3C Recommendation and has features in flexibility and extensibility.

The case studies in this research, particularly relating to how the three contribute to the different challenges found in building digital library initiatives, provide some insights which developers might take into account, and which could serve as valuable practical experience for improving or starting their digital libraries. A study of these three initiatives and my research in general uncovered a number of recommendations for future work which are detailed in Section 3 of this Chapter. This study also demonstrates that libraries will have to change in fundamental ways to survive and evolve in the digital age.

10.2 Looking to the Future

As of 2005, there are two important projects being undertaken in the Library of Congress which deserve full attention over the coming years. The first project is the National Digital Information and Infrastructure Preservation Program, which was discussed in Section 2.2 of Chapter 3 and Section 3.3 of Chapter 6; the second project is the LC Digital Audio-Visual Preservation Prototyping Project, which was discussed in several sections of Chapters 8 and Section 3.1.2 of Chapter 9. These two projects are deploying existing new initiatives, including XML and METS, which are playing a key role in tackling the sophisticated challenges. The Library of Congress is a large institution that reacts relatively slowly to rapidly changing technology, but it is in a position to recommend standards once projects are mature. Furthermore, the projects demonstrate examples of partnership not only in the digital library arena but also in the wider world of information technology.

The development of the digital library is a trend which combines social, economic, legal and technological advances. As I have shown, XML has been developed and has taken advantage of other technologies that together have made possible the World Wide Web. The application of XML in digital libraries is still in its early stages. The development of the Web from HTML to XML is benefiting industries involved in presenting data on the Web, and digital libraries are no exception. Digital libraries are more complex than many commercial applications that use XML because their data are less structured, and so the impact of XML in digital libraries is taking longer to reveal itself, even though the use of XML in digital libraries had a head start over its use in other applications such as commercial ones. From my research, I have noticed that much experimentation is taking place and XML is beginning to play a crucial role in digital library development.

Towards semantic digital libraries

As I discussed in Section 1.1.1 of Chapter 9, "inserting markup in a text is an act of interpretation". The problem is how descriptive markup conveys meaning, and how the technologies support the expression and process the meaning. Both De Belder [1993] and Hockey [2000, Chapter 5] considered it was much easier and more sensible to put the intelligence into the text in the form of encoding than it is to build sophisticated intelligence into computer programs. XML supports a document's meaningful structure but does not explicitly represent fundamental semantic relationships among document components and features. Cover [1998] thought XML itself offers only a standard for the syntax of an unspecified markup language. XML can only help humans predict what information might lie "between the tags", and this information, of course, can only be predicted by human intelligence understanding tags which depend upon natural, usually English, language.

Markup requires consensus on meaning in an area where there has been no need to achieve this in the past. A traditional catalogue entry uses natural language which can be understood by the user, assuming that that user is competent in that natural language. 'Composed by', 'edited by', 'played by' all indicate different kinds of creatorship which can be understood by the experienced user of a catalogue. How can this be incorporated in a standard way in markup? This is the core of the problem. To add value using markup is only of use if the value being added is understood by the retriever, and, for this to be achieved, the added value must be applied in a consistent way. Librarians have developed MARC based on a sophisticated set of rules which were, in turn, based on the Anglo-American Cataloguing Rules, which have been refined over many years. Digital librarians have sought to make the creation of metadata in the digital environment an easy activity, so that it can be achieved in quantity if not in quality. The creators of digital resources can provide metadata and relieve the digital librarians of this task. Dublin Core has been devised as a tool to

make this possible, or at least economically achievable. However, the result, if it even achieves quantity, does not achieve the level of quality necessary to facilitate the retrieval of the appropriate resources from the global digital library. The TEI Header allows the encoder to specify his or her interpretation of a particular tag; yet, semantic agreement probably needs something not just as precisely defined as the traditional catalogue entry but more precisely defined in order to achieve a semantically efficient digital library.

Although there have been related efforts in trying to address the semantics for XML markup by developing standards and recommendations such as the W3C Schema, ISO Topic Maps, RDF and the W3C Semantic Web, more research on the Semantic Web, which is taking a different view from that of Berners-Lee et al. [2001] at W3C, is beginning to emerge [Renear et al., 2002; Renear et al., 2003]. As Hockey [2000, Chapter 10] pointed out in her conclusions, one of the solutions to closing the gap between researchers and programmers is that scholars from all disciplines should involve themselves at all possible stages in the development. Perhaps the efforts of the Renear group are a positive inspiration and a meaningful start to a development towards a semantic digital library.

10.3 Recommendations for Digital Library Development

We have seen that XML has been implemented in a wide range of library operations as research projects or to improve their services [Miller, 2002; Banerjee, 2002; DEF, 2005; Lee, 2005]. However, there is not yet, in practice, any real XML-based digital library operation that has pulled the various application projects together in a way that allows users to access the various content and services better through a consistent mechanism across collections and digital libraries. Therefore, a great effort should be made in research initiatives and existing practices need to be taken into account to investigate the implementation of XML in digital libraries. A number of recommendations became evident during this research, and I summarize them below.

1. Automatic tagging is an area where a great deal more research needs to be done. The information industry, government agencies, research institutions and the library community could work together and pool their resources to improve automatic tagging. As discussed in Chapter 8, the creation and management of markup in the transcription process can be costly and laborious, automatic tagging has been a major concern for librarians in marking up library materials. We have seen two of my case study libraries, Michigan and Perseus, applying this technique as part of their digital library workflows. Standardized metadata with XML structures facilitates programs

which automatically identify structural features. Basic structural conversion can be done automatically. However, interpretative markup or markup that cannot be rule-based will need manual input by skilled workers or domain experts who understand the content. Intelligent automatic markup still has limitations in its reasoning power, which could leave intellectual content to specialists.

2. Digital libraries are applying an increasingly sophisticated use of information technology in their digital library systems. Such libraries need more computing professionals to manage the systems, especially people with XML knowledge. We have seen through our investigations reported in Chapter 9 that it would be helpful if librarians had some knowledge of XML, as XML will be increasingly involved in the whole process of library operations. Librarians in the digital age will need to blur the distinctions between library and computing professions and, as far as librarians are concerned, change their role from that of traditional book-keepers to those of "cybrarian", working closely with computing personnel. This could also apply to the scenario of Perseus where, as I pointed out in Chapter 9, Perseus management raised the issue of the dearth of institutional structures in place to produce corpus editors, who know enough about both computational algorithms and the documents in the corpus to adapt these techniques for the corpus in question. Digital libraries are looking for staff who are forward-looking (that is, visionaries), and can particularly act as team partners to explore and advise institutions on key issues in developing digital libraries. More attention must therefore be paid to the retention of staff with XML knowledge, which is valued in the commercial world as well as in academic digital library development.

3. METS promises to be able to provide an XML-based integrated structure in the digital library by pulling together the XML aware metadata standards and initiatives. The infrastructure is now in place. As discussed in Chapter 8, the follow-up task is to establish common schemes for the naming of digital objects in order to link these schemes to protocols for object transmission, metadata and object type classifications. The achievement of consensus in naming schemas for digital objects will allow a unique global reference to be employed, and facilitate resource sharing, linkages and interoperation among digital library systems. The long-term challenge is collaboration and coordination in the library community to fulfil the difficult task I discussed above.

4. As we have seen in Chapters 2, 8 and 9, XML exists in the company of a wealth of free tools from active user communities such as Linux, Java technology, Apache, MySQL and Perl language, which make possible a scalable, distributed digital library platform at less cost than commercial ones. These tools reduce costs, lowering the barriers to innovation, which in turn helps foster digital library development. I discussed in Chapter 8 that Perseus is entirely in the open source environment; Michigan's technologies are working in the direction of open source software as requested by their DLXS members. It would be beneficial if the industry would

provide more tools which could be used with open source software (as they have for other software such as Java) enhancing XML's position in the open source movement. Digital libraries should then consider more infrastructure based on XML-friendly open source software as a long-term management priority.

5. As we have seen in Chapter 2, XML is likely to be used increasingly in the future for electronic journal and database publishing. It would be useful to see if XML data fares better than other formats in the creation of effective indexes which will make it easier for researchers and students to use these databases. Libraries, especially academic libraries, tend to connect to numerous primary resource databases in response to the requirements of researchers and students. Each of the databases has its own search screen and special search features. This makes it difficult for the users, who need to familiarize themselves with many different search systems. So, this unfriendly search environment requires a major investment, which is a drain on the resources of library and user alike. More work needs to be done on developing a cross-database search interface to those resources which are encoded in XML.

6. Support for the XML Schema is being implemented for the major XML software offerings. As discussed in Chapter 8, new XML-based initiatives are being developed directly with XML Schemas such as METS and MODS. The developer can process schema definitions with standard XML tools and services such as DOM, SAX, XPath, and XSLT. In the future, it seems that more documents in digital libraries will be converted from SGML to XML, and so the XML Schema will be used more as time goes on. This would be a crucial point for heterogeneous digital library environments. It would be helpful if digital libraries invested effort in implementing schemas, an XML-language constraint language, as a priority, so as to ensure the success of their systems. Since TEI is the main metadata system in my three case studies, and also in many other digital projects, in order to meet the needs of more complicated applications in digital libraries, libraries should monitor the way in which TEI is adopting XML Schema language.

7. We have seen in Chapter 2 that there is a limited number of standards and minimal support from industry when it comes to multimedia content. Compared with XML-based text content, technology for XML-based non-text has made slower progress, partially because there has not been sufficient bandwidth to transfer these data across the Internet though this situation changed around 2005 in the experience of the author when institutional and domestic bandwidth was increase. It would be beneficial if more research institutions and industry jointly contribute to non-text content initiatives, especially those using XML-based technology. As we have seen, the Perseus multimedia digital library has shown that XML can be used fruitfully in teaching and learning although, as discussed in Chapter 9, there is still room for improvement: a global digital library requires a high-band-width network to deliver rich, interactive content and applications and services to users. I believe that the global delivery of multimedia resources is an essential part of

the mission of the digital library in providing a universally available resource for teaching and learning in life-long education. Only if these components come together will the digital library be a realistic goal.

8. As the dynamics of information and communication technologies have contributed to the value of information at all levels of society, particularly in education, digital libraries could invest more effort in XML-based information management projects, applying the XML namespaces and linking ability discussed in Chapter 2. This approach would allow applications to reuse existing languages instead of writing new ones from scratch. XML linking provides more functionality than the simple and single HTML hypertext link. XML namespaces and XML smart linking contribute to document integration in the digital environment because they provide a base for dynamic information management to gather and organize relevant internal and global information sources. This can also work alongside the XML-based Semantic Web and Topic Maps discussed in Chapter 5. The two technologies describe knowledge structures and associate them with information resources. From the users' point of view, it would help if libraries shifted from the traditional static way of library service to repackaging information resources and making these applicable to users.

9. Digital libraries need to monitor closely the development of the Web Services activities at W3C and implement the technologies where appropriate. As we have seen in Chapter 8, this technology is increasingly attracting the attention of the library and information communities. Web Services use XML as the file format and this leads to the use of emerging Web services technologies such as SOAP, either as an internal or external service available over the Internet. These services, which can be connected together to create the information technology systems of the future, will require less customized software in organizations and more creativity in the connections between the services.

10. The digital library combines many disciplines which require new evaluation methods to facilitate research into the full range of digital library issues. If digital libraries seriously invested more effort and cooperated with the relevant research groups to build a collection of evaluation methods and techniques that could be used within the framework of digital library evaluation, they would be proceeding in the right direction. Different evaluation methods may be used for different purposes and levels of analysis. The evaluation should be an ongoing and long-term process. As I remarked in Chapter 9, given the shortage of evaluation schemes that the three institutions in case studies had applied in their digital libraries, I would suggest that they put more effort in this area.

11. Digital libraries ought to be incorporated into their parent institutions or in the case of independent digital libraries at least be set on an institutional footing and not rely too heavily on particular individuals or funds. In addition, the collections would benefit from an institutional base and ready market. Also, staff, technology and other resources would be shared. If the digital

library is incorporated into an institution's library, library staff would have close contact with the digital library, which will benefit their skills and make them better prepared for the digital age. As we have seen in Chapter 9, two of my case study libraries, Michigan and the Library of Congress, are examples of this.

12. Digital libraries are moving from project status, where experimentation and the encouragement of innovation and development are the most significant features, to a fully fledged service, which requires stability and cohesion. There are a number of issues that need to be taken into account. The most important are maintaining the quality of the data and flexibility. To achieve this, it is important to keep up with the standards and recommendations that are being set for the Web. The World Wide Web is still very young and advances continue to make development much easier. Continually changing technology is a vital issue. As discussed in Chapters 8 and 9, my three case study libraries invested their data with standards and recommendations, so would have had fewer problems of migration.

13. Finally, digital libraries could best succeed with strong partnerships and collaboration. Such initiatives are not trivial tasks and would demand a considerable amount of time and resources from the participants. In this sense, digital libraries need a stable funding stream that allows them to continue to make valuable contributions to more initiatives. As I pointed out in Chapter 9, it is crucial that governments, industry, private enterprises and research institutions devote research and development funding to XML and digital libraries, to enable a wealth of experience to be gained to strengthen and further the state-of-the-art in XML-based digital libraries.

10.4 Conclusion

The digital library has many of the qualities of a traditional library, with the added benefit that a digital library, with its connection to the Internet, becomes a global library - the "library without walls". It represents a vision of the library extending beyond physical boundaries and reaching out to the world at large. The Google project discussed in Chapter 3 gives evidence of this concept. Vast digital libraries of information are becoming available on the Internet, as the result of emerging technologies for the processing of data in different media.

Descriptive markup is seen as a superior tool for document portability and for building up digital library collections, and this is being realized through the availability and extensive use of XML. Descriptive markup qualifies text and has made it possible to deliver large texts in small pieces. Programs can operate on that text in a variety of ways; it becomes easier to use many different

kinds of software with the same machine-readable text. As I concluded in Chapter 2, XML was designed to be easily available on the Web (unlike SGML). This is the key aspect that SGML cannot but XML does provide. Also, as I concluded in Chapter 8, XML has a high impact on metadata and interoperability in digital library development, and the two instances are key components in building global networked digital libraries.

As I have indicated throughout the thesis, more Internet and library systems standards are being based on XML as time goes on. XML associated initiatives, software and tools will become more available and user-friendly over time and are likely to be implemented by institutions. In the future, library services will become more incorporated into network and Internet technology, and XML is the new technology for the exchange of data for the Web. So, the impact of XML on digital library developments is sure to increase. Digital library projects will suffer if they do not seriously consider XML. This can be seen in my three case study digital libraries: Perseus, a technology-oriented digital library project; Michigan, an academic digital library providing part of an integrated hybrid library service to users; LC, a national library which is based in a traditional reference library, are all heading towards the same direction, that is incorporating more XML technology in their digital library development as time goes on.

Bibliography:

Abate, Tom. (1997). Publishing Scientific Journals Online. *BioScience*, 47(3). Online. Available: http://www.aibs.org/bioscience/bioscience-archive/vol47/Mar97abate.html (Accessed on 5 February 2004)

Abitboul, Serge. (1997). Querying Semi-Structured Data. In: Foto N. Afrati and Phokion G. Kolaitis, editors, *Proceedings of the 6th International Conference on Database Theory (ICDT'97), Delphi, Greece, 8-10 January 1997.* Berlin; Heidelberg: Springer-Verlag, pp.1-18.

Abiteboul, Serge et al. (1997). The Lorel Query Language for Semistructured Data. *International Journal on Digital Libraries,* 1(1): 68-88.

ACM. (n.d.). *ACM International Conference on Digital Libraries.* Online. Available: http://www.acm.org/ (Accessed on 6 January 2004)

ACM DL. (1999). *The ACM Digital Libraries Conference (DL '99), 11-14 August, Berkeley, California, USA.*

AHDS. (2000). *Creating Digital Resources for the Visual Arts: Standards and Good Practice, Section 3.1 Introducing Digital Images and Image Creation.* Online. Available: http://vads.ahds.ac.uk/guides/creating_guide/sect31.html (Accessed on 28 February 2004)

AHDS. (2003). *AHDS Guides to Good Practice.* Online. Available: http://ahds.ac.uk/litlangling/creating/guides/index.htm (Accessed on 1 December 2003)

ALA. (2003). *Top Technology Trends by Topic.* Online. Available: http://www.ala.org/ala/lita/litaresources/toptechtrends/toptechtrends.htm (Accessed on 8 September 2004)

ALCTS Continuing Education Task Force (Action Item 5.3). (2003). *Cataloging for the 21st Century: a Proposal for Continuing Education for Cataloging Professionals.* Online. Available: http://darkwing.uoregon.edu/~chixson/cetf/CETF_Final_Report.pdf (Accessed on 12 September 2004)

American Memory. (n.d.). *American Memory Collections: Other Collaborative Projects.* Online. Available: http://memory.loc.gov/cgi-bin/query/ S?ammem/collections:@field(OTHER+@band(Other)):heading=Other+Collaborative+Projects (Accessed on 16 November 2002)

Anderson, Martha. (1999). A Tool for Building Digital Libraries. *D-Lib Magazine*, 5(2). Online. Available: http://www.dlib.org/dlib/february99/02journalreview.html (Accessed on 27 January 2004)

Andresen, Leif. (2003). *Open Letter from Leif Andresen, Chair. Danish Standard 24 – Information and Documentation to Pat Harris ISO TC46/SC4 Secretariat. (ISO TC46*

SC4 n524). Online. Available: http://www.niso.org/international/SC4/n524.pdf (Accessed on 15 September 2003)

Apache Software Foundation. (2000). *Schema Implementation Limitations.* Online. Available: http://xml.apache.org/xerces-j/schema.html (Accessed on 24 January 2004)

Apache Software Foundation. (2005). *Apache Xindice.* Online. Available: http://xml.apache.org/xindice/ (Accessed on 13 August 2005)

Apps, Ann and MacIntyre, Ross. (2000). *XML: Using an Evolving Standard in Electronic Publishing.* Online. Available: http://epub.mimas.ac.uk/papers/appsmacep2000_full.html (Accessed on 25 July 2003)

Arbortext. (n.d.). *XML for Managers: Evaluating SGML vs. XML from a Manager's Prospective.* Online. Available: http://ora.kangnung.ac.kr/xml/xmlwp.html (Accessed on 5 December 2003)

ARL. (1995). *Definition and Purposes of a Digital Library.* Association of Research Libraries. Online. Available: http://www.ifla.org/documents/libraries/net/arl-dlib.txt (Accessed on 9 September 2001)

ARL. (2003). *Libraries Urge Justice Department to Block Cinven and Candover Purchase of Bertelsmannspringer: Publisher Mergers Threaten Access to Crucial Research.* Online. Available: http://www.arl.org/scomm/MergerRelease-530.pdf (Accessed on 6 February 2004)

Arms, Caroline R. (1996). Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress. *D-Lib Magazine*, April. Online. Available: http://www.dlib.org/dlib/april96/loc/04c-arms.html (Accessed on 1 November 2002)

Arms, Caroline R. (1999). Getting the Picture: Observations from the Library of Congress on Providing Online Access to Pictorial Images. *Library Trends*, 48(2): 379-409. Online. Available: http://memory.loc.gov/ammem/techdocs/libt1999/libt1999.html (Accessed on 15 January 2003)

Arms, Caroline R. (2000). Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program of the Library of Congress. *RLG DigiNews*, 4(3). Online. Available: http://www.rlg.org/preserv/diginews/diginews4-3.html#feature1 (Accessed on 7 October 2001)

Arms, Caroline R. (2004). *Requesting Update Information on NDL XML Activities.* E-mail to Naicheng Chang. 19 August 2004.

Arms, William Y. (1999). Preservation of Scientific Serials: Three Current Examples. *Journal of Electronic Publishing*, 5(2). Online. Available: http://www.press.umich.edu/jep/05-02/arms.html (Accessed on 1 September 2002)

Arms, William Y. (2000). *Digital Libraries*. Cambridge, Massachusetts: the MIT Press, pp.227-237.

Arms, William Y. (2005). Information Science as a Liberal Art. *Interlending & Document Supply*, 33(2): 81-84.

Arms, William Y et al. (1997). An Architecture for Information in Digital Libraries. *D-Lib Magazine*, February. Online. Available: http://www.dlib.org/dlib/february97/ cnri/02arms1.html (Accessed on 15 January 2003)

Atkins, Daniel E. (1996). The University of Michigan Digital Library Project: the Testbed. *D-Lib Magazine,* July/August. Online. Available: http://www.dlib.org/dlib/july96/ 07atkins.html (Accessed on 1 October 2002)

Baca, Murtha, editor. (1998). *Introduction to Metadata: Pathway to Digital Information.* Los Angeles, California: Getty Information Institute, pp.1-3.

Baird, H S et al., editors. (1992). *Structured Document Image Analysis*. Berlin: Springer-Verlag, pp.477-556.

Baker, Thomas. (2000). A Grammar of Dublin Core. *D-Lib Magazine,* 6(10). Online. Available: http://www.dlib.org/dlib/october00/baker/10baker.html (Accessed on 19 June 2001)

Baker, Thomas and Lynch, Clifford A. (1998). Summary Review of the Working Group on Metadata. In: Peter Schäuble and Alan F. Smeaton, editors, *Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research.* Joint NSF-EU Working Groups on Future Directions of Digital Libraries Research, 12 October 1998. Online. Available: http://www.dli2.nsf.gov/ internationalprojects/eu_d.html (Accessed on 24 November 2003)

Banerjee, Kyle. (2002). How Does XML Help Libraries? *Computers in Libraries,* 22(8). Online. Available: http://www.infotoday.com/ cilmag/sep02/ Banerjee.htm (Accessed on 14 October 2004)

Barry & Associates. (n.d.). *Object-Relational Mapping.* Online. Available: http://www.object-relational.com/ (Accessed on 14 June 2001)

Bater, Bob. (2004). Topic Maps: Indexing in 3-D. In: Alan Gilchrist and Barry Mahon, editors, *Information Architecture: Designing Information Environments for Purpose.* London: Facet Publishing, Chapter 8.

Beggs, Josh and Thede, Dylan. (2001). *Designing Web Audio.* Sebastopol, California: O'Reilly, pp. 210-220.

Bekaert, Jeroen and Van de Sompel, Herbert. (2005). A Standards-Based Solution for the Accurate Transfer of Digital Assets. *D-Lib Magazine,* 11(6). Online. Available: http://www.dlib.org/dlib/june05/bekaert/06bekaert.html (Accessed on 10 July 2005)

Berbecaru, Diana et al. (2000). Towards Concrete Application of Electronic Signature. In: *Proceedings Le Tecnologie dell'Informazione e della Comunicazione Come Motore di Sviluppo del Paese (AICA Annual Conference), Taormina, Italy, 27-30 September 2000.* pp.543-561.

Bergerud, Marly. (1987). *Word and Information Processing: Concepts of Office Automation*, 3rd edition, New York; Chichester: Wiley, p.176.

Bergman, Michael K. (2001). The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1). Online. Available: http://www.press.umich.edu/jep/07-01/bergman.html (Accessed on 5 September 2001)

Berners-Lee, Tim. (1989). *Information Management: a Proposal, In-House Technical Document, CERN. (revised 1990 with Robert Cailliau).* Online. Available: http://www.w3.org/History/1989/proposal.html (Accessed on 25 January 2004)

Berners-Lee, Tim. (1997). *Metadata Architecture.* Online. Available: http://www.w3.org/DesignIssues/Metadata (Accessed on 25 October 2000)

Berners-Lee, Tim. (1998). *Semantic Web Road Map.* Online. Available: http://www.w3.org/DesignIssues/Semantic.html (Accessed on 11 October 2001)

Berners-Lee, Tim and Swick, Ralph R. (1999). *Frequently Asked Questions about RDF.* W3C Technology and Society Domain. Online. Available: http://www.w3.org/RDF/FAQ (Accessed on 22 October 2000)

Berners-Lee, Tim and the W3C Team. (1997). *W3C Data Formats, W3C NOTE 29 October 1997.* Online. Available: http://www.w3.org/TR/NOTE-rdfarch (Accessed on 25 October 2000)

Berners-Lee, Tim et al. (1999). *Web Architecture: Describing and Exchanging Data, W3C Note 7 June 1999.* Online. Available: http://www.w3.org/1999/06/07-WebData (Accessed on 25 October 2000)

Berners-Lee, Tim et al. (2001). The Semantic Web: a New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Feature Article in *Scientific American*, May. Online. Available: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 (Accessed on 11 October 2001)

Biezunski, Michel and Newcomb, Steven R. (2001). XML Topic Maps: Finding Aids for the Web. *IEEE Multimedia*, 8(2): 104-108.

Billington, James. (1995a). *The Mission and Strategic Priorities of the Library of Congress.* Online. Available: http://lcweb.loc.gov/ndl/mission.html (Accessed on 5 November 2002)

Billington, James. (1995b). Preserving the Inclusive Nature of American Public Libraries. *National Digital Library Periodical Report,* 5. A Conference Speech with Topic: the Transformation of the Public Library: Access to Digital Information in a Networked World, the Library of Congress, 8 December 1995. Online. Available: http://www.loc.gov/ndl/jan-feb.html (Accessed on 9 November 2002)

Blandford, Ann. (2004). *Understanding Users' Experiences: Evaluation of Digital Libraries.* Workshop Report at DELOS Evaluation Workshop. Online. Available: http://www.uclic.ucl.ac.uk/annb/DLUsability/DELOSabstract.pdf (Accessed on 20 March 2005)

Blandford, Ann and Buchanan, George. (2002). *Usability of Digital Libraries.* Workshop Report at JCDL' 02. Online. Available: http://www.uclic.ucl.ac.uk/annb/DLUsability/SIGIR.pdf (Accessed on 20 February 2004)

Bonifati, Angela and Ceri, Stefano. (2000). Comparative Analysis of Five XML Query Languages. *SIGMOD RECORD*, 29(1): 74-77.

Bonn, María S et al. (1999). A Report on the PEAK Experiment. *D-Lib Magazine*, 5(6). Online. Available: http://www.dlib.org/dlib/june99/06bonn.html (Accessed on 28 October 2002)

Booth, David et al., editors. (2003). *Web Services Architecture.* Online. Available: http://www.w3.org/TR/2003/WD-ws-arch-20030808/ (Accessed on 3 January 2004)

Borgman, Christine L et al. (2000). Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: a Case Study of the Alexandria Digital Earth Prototype (ADEPT). *Library Trends*, 49(2): 228-250.

Bos, Burt. (2005). *Web Style Sheets Home Page.* Online. Available: http://www.w3c.org/Style/ (Accessed on 12 August 2005)

Bosak, Jon. (1996). *DSSSL Online Application Profile.* Online. Available: http://www.ibiblio.org/pub/sun-info/standards/dsssl/dssslo/do960816.htm (Accessed on 30 July 2005)

Bosak, Jon. (1997). *XML, Java, and the Future of the Web.* Online. Available: http://www.ibiblio.org/pub/sun-info/standards/xml/why/xmlapps.htm (Accessed on 28 May 2001)

Bourret, Ronald. (2001). *Mapping DTDs to Databases.* Online. Available: http://www.xml.com/pub/a/2001/05/09/dtdtodbs.html (Accessed on 15 June 2001)

Bourret, Ronald. (2004). *XML and Databases.* Online. Available: http://www.rpbourret.com/xml/XMLAndDatabases.htm#nativedb (Accessed on 16 August 2005)

Bray, Tim et al., editors. (1998). *Extensible Markup Language (XML) 1.0, W3C Recommendation 10 February 1998.* Online. Available: http://www.w3.org/TR/1998/REC-xml-19980210 (Accessed on 30 September 2000)

Bray, Tim et al., editors. (1999). *Namespaces in XML, W3C Recommendation 14 January 1999.* Online. Available: http://www.w3.org/TR/REC-xml-names/ (Accessed on 24 December 2000)

Bray, Tim et al., editors. (2004). *Extensible Markup Language (XML) 1.0 (Third Edition), W3C Recommendation 4 February 2004.* Online. Available:

http://www.w3.org/TR/2004/REC-xml-20040204 (Accessed on 24 May 2004)

British Library. (2003). *The British Library's Co-operation and Partnership Programme (CPP)*, Online. Available: http://www.bl.uk/concord/concord.html (Accessed on 6 January 2004)

Brophy, Peter and Wynne, Peter M. (1997). *Management Information Systems and Performance Measurement for the Electronic Library*. eLib Supporting Study (MIEL2) Final Report. Online. Available: http://www.ukoln.ac.uk/services/elib/papers/ supporting/pdf/mis.pdf (Accessed on 11 September 2001)

Broughton, Vanda. (2001). Faceted Classification as a Basis for Knowledge Organization in a Digital Environment: the Bliss Bibliographic Classification as a Model for Vocabulary Management and the Creation of Multidimensional Knowledge Structures. *The New Review of Hypermedia and Multimedia*, 7: 67-102.

Brown, Alex. (2001). *An Introduction to Unicode's Role in XML.* Online. Available: http://www.griffinbrown.co.uk/griffin-brown-unicode-article.pdf (Accessed on 15 November 2003)

Bryan, Martin. (1992). *An Introduction to the Standard Generalized Markup Language (SGML).* Online. Available: http://www.oasis-open.org/cover/bryanIntro1992.html (Accessed on 31 May 2001)

Bryan, Martin. (1997). *SGML and HTML Explained.* Harlow: Addison Wesley Longman, pp.224-226.

BUILDER. (2001). *Builder Final Report.* Online. Available: http://builder.bham.ac.uk/finalreport/pdf/fr.pdf (Accessed on 3 December 2003)

Bunker, Geri and Zick, Greg. (1999). Collaboration as a Key to Digital Library Development: High Performance Image Management at the University of Washington. *D-Lib Magazine*, 5(3). Online. Available: http://www.dlib.org/dlib/march99/bunker/03bunker.html (Accessed on 15 July 2003)

Burnard, Lou. (1995). *Text Encoding for Information Interchange: an Introduction to the Text Encoding Initiative.* Online. Available: http://www.tei-c.org/Vault/SC/J31/ (Accessed on 26 September 2001)

Burnard, Lou. (2002). *Updated PizzaChef Tool Supports XML DTD Generation.* OASIS News Cover Stories. Online. Available: http://xml.coverpages.org/ni2002-02-19-a.html (Accessed on 30 September 2003)

Burnard, Lou and Light, Richard. (1996). *Three SGML Metadata Formats: TEI, EAD,and CIMI: a Study for BIBLINK Work Package 1.1.* Online. Available: http://hosted.ukoln.ac.uk/biblink/wp1/sgml/ (Accessed on 17 September 2003)

Bush, Vannevar. (1945). As We May Think. *Atlantic Monthly,* July, pp.101-108.

Calanag, María Luisa et al. (2002). Linking Collection Management Policy to Metadata for

Preservation: a Guidance Model to Define Metadata Description Levels in Digital Archives. In: *Proceedings of the Internaltional Conference on Dublin Core and Metadata for e-Communities 2002, Florence, Italy, 13-17 October 2002.* Florence: Firenze University Press, pp.35-43.

Campbell, Laura. (1995). How Do You Staff the Huge Task of Digitizing a Culture by the Year 2000? *National Digital Library Periodical Report,* 4. Online. Available: http://lcweb.loc.gov/ndl/nov-dec.html (Accessed on 30 Janurary 2003)

Caplan, Priscilla. (2000). *International Metadata Initiatives: Lessons in Bibliographic Control.* Paper presented in the Proceedings of the Bicentennial Conference on Bibliographic Control in the New Millenium: Confronting the Challenges of the Networked Resources and the Web, the Library of Congress Cataloguing Directorate, 15-17 November 2000. Online. Available: http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html (Accessed on 2 October 2003)

Carr, Reg. (2001). *Towards the Hybrid Library: the National Perspective in the UK.* Paper presented in the MALIBU Conference, King's College London, 26 March 2001. Online. Available:http://www.bodley.ox.ac.uk/librarian/malibu2001/malibu2001.htm (Accessed on 3 January 2004)

Carvalho, De Joaquim and Cordeiro, María Inês. (2002). *XML and Bibliographic Data: the TVS (Transport, Validation and Services) Model.* Paper presented in the 68th IFLA Council and General Conference, Glasgow, 18-24 August 2002. Online. Available: http://www.ifla.org/IV/ifla68/papers/075-095e.pdf (Accessed on 20 September 2003)

Cattell, R G G. (1992). *Object Data Management: Object-Oriented and Extended Relational Database Systems.* Reprinted with corrections. Reading, Massachusetts: Addison-Wesley. passim.

Cattell, R G G, editor. (1996). *The Object Database Standard: ODMG-9.* San Francisco, California: Morgan Kaufmann, Chapter 1.

Cattell, R G G and Barry, Douglas K, editors. (2000). *The Object Data Standard: ODMG 3.0.* San Francisco; London: Morgan Kaufmann, pp. 4-5.

CDL. (1999). *Digital Image Collection Standards.* Online. Available: http://www.cdlib.org/inside/groups/stas/standards/ (Accessed on 26 December 2003)

CDL. (2001). *Digital Objects Standard: Metadata, Content and Encoding.* Online. Available: http://www.cdlib.org/news/pdf/CDLObjectStd-2001.pdf (Accessed on 26 December 2003)

CEDARS. (n.d.). Online. Available: http://www.leeds.ac.uk/cedars/ (Accessed on 14 July 2002)

Chachra, Krisha. (2002). *VTLS Inc. Announces FRBR Implementation: VIRTUA ILS Now Supports FRBR.* Online. Available: http://www.vtls.com/Corporate/Releases/2002/20020514b.shtml (Accessed on 12 November 2004)

Chamberlin, Don et al., editors. (2003). *XML Query (XQuery) Requirements, W3C Working Draft 2 May 2003.* Online. Available: http://www.w3.org/TR/xquery-requirements/ (Accessed on 11 June 2003)

Chang, Naicheng and Perng, Jinhuai. (2001). Information Search Habits of Graduate Students at Tatung University. *The International Information and Library Review,* 33(4): 341-346.

Chapman, Stephen and Kenny, Anne R. (1996). Digital Conversion of Research Library Materials: a Case for Full Information Capture. *D-Lib Magazine*, October. Online. Available: http://www.dlib.org/dlib/october96/cornell/10chapman.html (Accessed on 22 November 2002)

Chen, Minder et al. (2003). The Implications and Impacts of Web Services to E-commerce Research and Practices. *Journal of Electronic Commerce Research,* 4(4): 128-139.

CHLT. (n.d.). *Cultural Heritage Language Technologies*. Online. Available: http://www.chlt.org/ (Accessed on 15 September 2004)

Chodorow, S and Lyman, P. (1998). The Responsibilities and Universities in the New Information Environment. In: B. L. Hawkins and P. Battin, editors, *the Mirage of Continuity: Reconfiguring Academic Information Resources for the 21st century.* Washington, DC: Council on Library and Information Resources and Association of American Universities, pp.61-78.

CILIP. (2004). *What You Need to Know about Metadata.* CILIP Training Workshops, 27 October 2004, London. Online. Available: http://www.cilip.org.uk/training/training/ ict/metadata.htm (Accessed on 12 September 2004)

CIMI. (2003). *CIMI XML Testbed.* Online. Available: http://www.cimi.org/wg/ xml_spectrum/xml_CFP.html (Accessed on 4 January 2004)

Clarke, Kevin S. (2001). *Medlane/XMLMARC Update: from MARC to XML Database.* Paper presented in the Medical Library Association (MLA) National Conference, 25-30 May 2001, Orlando, Florida. Online. Available: http://xmlmarc.stanford.edu/ MLA2001/medlane.html (Accessed on 15 July 2001)

Clarke, Kevin S. (2002). Updating MARC Records with XMLMARC. In: Tennant, Roy, editor, *XML in Libraries*. New York: Neal-Schuman, pp.3-16.

CLIC. (1996). *CLIC: an Electronic Version of Chemical Communications.* Online. Available: http://www.ch.ic.ac.uk/rzepa/watoc96/wa_8.html (Accessed on 11 July 2001)

CLIR. (2001). *Building and Sustaining Digital Collections: Models for Libraries and Museums.* Online. Available: http://www.clir.org/pubs/reports/ pub100/pub100.pdf (Accessed on 15 June 2004)

CNI. (1998). *CNI Spring 1998 Task Force Meeting Final Report.* Online. Available: http://www.cni.org/tfms/1998a.spring/final_report98Stf.html (Accessed on 8 October 2001)

Cole, Timothy W et al. (2000). *Using XML, XSLT, and CSS in a Digital Library*. Paper presented in the ASIS 2000: Knowledge Innovations, 63rd American Society for Information Science Annual Meeting, Chicago, Illinois, 15 November 2000. Online. Available: http://dli.grainger.uiuc.edu/idli/whitepapers/xmltechnologies.pdf (Accessed on 1 December 2003)

Columbia University Digital Library Project. (1998). *A Relational Model for MARC Bibliographic Data.* Online. Available: http://www.columbia.edu/cu/libraries/ inside/projects/metadata/model/whitepaper.html (Accessed on 10 July 2003)

Conway, Alan. (1993). Page Grammars and Page Parsing: a Syntactic Approach to Document Layout Recognition. In: *Proceeding of the 2nd International Conference on Document Analysis and Recognition, Tsukuba Science City, 20-22 October 1993.* Los Alamitos, California: IEEE Computer Society Press, pp.761-764.

Coombs, James H et al. (1987). Markup Systems and the Future of Scholarly Text Processing. *Communications of the Association for Computing Machinery*, 30(11): 933-947. Online. Available: http://www.oasis-open.org/cover/coombs.html (Accessed on 20 November 2001)

CORC. (1999). *CORC Project Participants Hold First Meeting, Dublin, Ohio, 24 May 1999.* Online. Available: http://www.oclc.org/research/publications/ archive/releases/ 1999-05-24.htm (Accessed on 28 October 2003)

Cornell Institute for Digital Collections. (n.d.). Online. Available: http://cidc.library.cornell.edu/ xml/ (Accessed on 8 January 2004)

COUNTER. (2004). *XML DTD Now Available for COUNTER Usage Reports.* Online. Available: http://www.projectcounter.org/xml.html (Accessed on 14 August 2004)

COUNTER. (2005). News & Actitivies. Online. Available: http://www.projectcounter.org/ news.html (Accessed on 12 August 2005)

COVAX. (2001). Online. Available: http://www.cultivate-int.org/issue3/covax/ (Accessed on 14 November 2001)

Cover, Robin. (1998). *XML and Semantic Transparency.* Online. Available: http://xml.coverpages.org/xmlAndSemantics.html (Accessed on 27 November 2003)

Cover, Robin. (2002). *SGML/XML Applications in Cross-Domains and Multi-Disciplinary Enterprises.* Online. Available: http://xml.coverpages.org/gen-apps.html (Accessed on 2 February 2004)

Crane, Gregory. (1998). The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine,* January. Online. Available: http://www.dlib.org/dlib/january98/01crane.html (Accessed on 27 October 2002)

Crane, Gregory. (2000a). Designing Documents to Enhance the Performance of Digital

Libraries: Time, Space, People and a Digital Library of London. *D-Lib Magazine*, 6(7/8). Online. Available: http://www.dlib.org/dlib/july00/crane/07crane.html (Accessed on 7 November 2002)

Crane, Gregory. (2000b). Extending a Digital Library: Beginning a Roman Perseus. *New England Classical Journal,* 27(3): 140-160.

Crane, Gregory. (2000c). From Greece to Rome: Building a Roman Perseus. *Journal for the Association of Classical Teachers*. Online. Available: http://www.perseus.tufts.edu/Articles/jact2000.html (Accessed on 24 March 2003)

Crane, Gregory and Rydberg-Cox, Jeffrey. (2000). New Technology and New Roles: the Need for "Corpus Editors". In: *Proceedings of the 5th ACM Conference on Digital Libraries, San Antonio, Texas, 2-7 June 2000*. New York: ACM Press, pp. 252-253.

Crane, Gregory et al. (1998). *A Digital Library for the Humanities: a Proposal Submitted to the NEH-NSF Digital Libraries Initiative, Phase 2.* Online. Available: http://tantalos.perseus.tufts.edu/Props/DLI2/dli2.html (Accessed on 29 October 2002)

Crane, Gregory et al. (2001). Building a Hypertextual Digital Library in the Humanities: a Case Study on London. In: *International Conference on Digital Libraries Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01), 24-28 June 2001, Roanoke, Virginia.* New York: ACM Press, pp. 426-434.

CrossRef. (2003). Online. Available: http://www.crossref.org (Accessed on 24 July 2003)

Dale, Robin. (2002). *CEDARS Final Workship.* Manchester, UMIST.

Day, Michael. (1997). Extending Metadata for Digital Preservation. *Ariadne*, 9. Online. Available: http://www.ariadne.ac.uk/issue9/metadata/ (Accessed on 24 November 2003)

Day, Michael. (1998a). *Issues and Approaches to Preservation Metadata.* Paper presented in the Joint RLG and NPO Preservation Conference: Guidelines for Digital Imaging, University of Warwick, 28-30 September 1998. Online. Available: http://www.rlg.org/preserv/joint/day.html (Accessed on 8 October 2001)

Day, Michael. (1998b). *Metadata for Preservation.* Online. Available: http://www.ukoln.ac.uk/metadata/cedars/AIW01.html (Accessed on 8 October 2001)

Day, Michael. (1999). *The Metadata Challenge for Libraries: a View from Europe.* Online. Available: http://www.ukoln.ac.uk/metadata/presentations/metadiversity/challenge.html (Accessed on 21 November 2001)

Day, Michael and Powell, Andy. (n.d.). *What is Metadata?* Online. Available: http://www.ukoln.ac.uk/metadata/ (Accessed on 8 October 2000)

DCQ. (2000). Online. Available: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/ (Accessed on 19 September 2003)

De Belder, Kurt. (1993). Electronic Texts: a Promise for Humanities Research. *Academic Computing and Networking at NYU,* 3(4). Online. Available:

http://library.nyu.edu/research/french/etext.html (Accessed on 6 December 2002)

DEF. (2005). *DEF XML Web Services.* Online. Available: http://defxws.cvt.dk/ (Accessed on 13 August 2005)

Denenberg, Ray. (2002). *[Zing] Overview.* ZIG Meeting at OCLC, April 2002. Online. Available: http://www.loc.gov/z3950/agency/zig/meetings/oclc2002/ppts/zing.ppt (Accessed on 27 October 2003)

DeRose, Steven. (1997). *The SGML FAQ Book: Understanding the Foundation of HTML and XML.* Boston: Kluwer Academic, pp.181-185.

DeRose, Steven. (1999a). What Do Those Weird XML Types Want, Anyway? In: Malcolm P. Atkinson et al., editors, *Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, 7-10 September, 1999.* San Francisco, California: Morgan Kaufmann, pp.721-724.

DeRose, Steven. (1999b). XML and TEI. *Computers and the Humanities,* 33: 11-30.

DeRose, Steven et al. (1998). *Queries on Links and Hierarchies.* Online. Available: http://www.w3.org/TandS/QL/QL98/pp/linkhier.html (Accessed on 27 January 2001)

DeRose, Steven et al. (2001). *XML Pointer Language (XPointer), Version 1.0.* Online. Available: http://www.w3.org/TR/2001/CR-xptr-20010911/ (Accessed on 27 December 2002)

DESIRE. (1999). Online. Available: http://desire.ukoln.ac.uk/registry/ (Accessed on 15 September 2003)

Deutsch, Alin et al. (1999). A Query Language for XML. *Computer Networks*, 31(11-16): 1155-1169.

Digimarc. (2000). *Digimarc Adopts XML for Digital Watermarking Technology.* Online. Available: http://www.digimarc.com/company/news/release.asp?newsID=137 (Accessed on 27 September 2003)

DLF. (1999). *TEI Text Encoding in Libraries: Guidelines for Best Encoding Practices, Version 1.0.* Online. Available: http://www.diglib.org/standards/tei.htm (Accessed on 7 November 2002)

DLF. (2005). *Digital Library Authentication and Authorization Architecture.* Online. Available: http://www.diglib.org/architectures/dcoverview.htm (Accessed on 15 April 2005)

DLI. (1999). Online. Available: http://www.dli2.nsf.gov/dlione/ (Accessed on 10 June 2002)

DLXS. (2003a). *Finding Aids, XSLT, METS Top List in 1st User Group Vote.* Online. Available: http://www.dlxs.org/about/news.html (Accessed on 15 November 2003)

DLXS. (2003b). *Institutional Contacts for Member Institutions.* Online. Available: http://www.dlxs.org/about/contacts.html (Accessed on 5 August 2005)

DOI. (2002). Online. Available: http://www.doi.org/ (Accessed on 15 August 2002)

Donnelly, Martin. (2003). *DigiCULT Technology Watch Briefing 7: the XML Family of Technologies.* Online. Available: http://www.digicult.info/downloads/

DigiCULT_TWB7_XML.pdf (Accessed on 5 November 2003)

DOREMI. (2004). DOREMI Research Group, Department of Computer Science, University of Helsinki. Online. Available: http://www.cs.helsinki.fi/group/doremi/ (Accessed on 21 October 2004)

Drewery, J O and Riley, J L. (1999). *International Broadcasting Convention (IBC 99).* Amsterdam, 10-14 September 1999. Conference Publication, pp. 90-95. Online. Available: http://www.bbc.co.uk/rd/pubs/papers/pdffiles/ibc99jod.pdf (Accessed on 3 March 2003)

Drucker, Johanna et al. (1999). *Refining Our Notions of What (Digital) Images Really Are.* A panel held at the ACH/ALLC International Humanities Computing Conference, Charlottesville, University of Virginia, 10 June 1999. Online. Available: http://www.iath.virginia.edu/ach-allc.99/proceedings/kirschenbaum.html (Accessed on 11 October 2001)

Dürst, Martin and Freytag, Asmus. (2003). *Unicode in XML and other Markup Languages, Unicode Technical Report, 20, W3C Note 13 June 2003.* Online. Available: http://www.w3.org/TR/2003/NOTE-unicode-xml-20030613/ (Accessed on 22 February 2004)

DYNIX. (2004). *Horizon Digital Library.* Online. Available: http://www.dynix.com/products/digital/ (Accessed on 28 January 2004)

Ebind. (n.d.). *Digital Page Imaging and SGML: an Introduction to the Electronic Binding DTD (Ebind).* Online. Available: http://sunsite.berkeley.edu/Ebind/ (Accessed on 23 February 2003)

eBooks.com. (n.d.). Online. Available: http://usa1.ebooks.com/ (Accessed on 9 February 2005)

Edinburgh Engineering Virtual Library. (2005). Online. Available: http://www.eevl.ac.uk/ (Accessed on 12 August 2005)

EDItEUR. (n.d.). Online. Available: http://www.editeur.org/ONIX%20International%20FAQ.html (Accessed on 30 January 2004)

eLibrary. (2002). *Online Libraries and Microcomputers: Ingenta Acquires HERON.* Online. Available: http://static.elibrary.com/o/onlinelibrariesandmicrocomputers/april012002/ingentaacquiresheron/index.html (Accessed on 1 March 2004)

EQUINOX. (2000). *EQUINOX: Library Performance Measurement and Quality Management System.* Online. Available: http://equinox.dcu.ie/ (Accessed on 3 February 2003)

Erickson, Janet C. (1997). *Options for Presentation of Multilingual Text: Use of the Unicode Standard.* Online. Available: http://xml.coverpages.org/ericksonUnicode.html (Accessed on 5 July 2005)

Etzioni, Oren and Weld, Daniel. (1999). *Automatic Reference Librarians for the World Wide Web.* Online. Available: http://www.cs.washington.edu/research/diglib/ (Accessed on 26

December 2001)

European Commission. (2002). *DigiCULT Full Report: Technological Landscapes for Tomorrow's Economy Cultural Unlocking the Value of Cultural Heritage.* Online. Available: http://www.salzburgresearch.at/fbi/digicult/ (Accessed on 13 February 2003)

Ex Libris. (n.d.). Online. Available: http://www.exlibris-usa.com/ (Accessed on 26 August 2003)

FEDORA. (n.d.). *The FEDORA Project: an Open-Source Digital Repository Management System.* Online. Available: http://www.fedora.info/ (Accessed on 3 January 2004)

Feizabadi, Shahrooz. (1998). History of the World Wide Web. In: Marc Abrams, editor, *World Wide Web – Beyond the Basics.* Upper Saddle River, NJ: Prentice Hall, Chapter 1. Online. Available: http://ei.cs.vt.edu/~wwwbtb/book/chap1/net_hist.html (Accessed on 2 July 2005)

Felstead, Alison. (2004). The Library Systems Market: a Digest of Current Literature. *Program: electronic library and information systems,* 38(2): 88-96.

Fernández, M et al. (1997). A Query Language for a Web-Site Management System. *SIGMOD Record,* 26(3): 4-11.

Ferraiolo, Jon et al., editors. (2003). *Scalable Vector Graphics (SVG) 1.1 Specification, W3C Recommendation 14 January 2003.* Online. Available: http://www.w3.org/TR/SVG11/ (Accessed on 19 February 2004)

FGDC. (n.d.). Online. Available: http://www.fgdc.gov/ (Accessed on 16 October 2001)

Fichter, Darlene. (2002). Migrating Native Law Cases from HTML to XML. In: Tennant, Roy, editor, *XML in Libraries.* New York: Neal-Schuman, pp.135-147.

Fleischhauer, Carl. (1995). American Memory Pilot: Seed of a Universally Available Library. *National Digital Library Periodical Report,* 4. Online. Available: http://lcweb.loc.gov/ndl/nov-dec.html (Accessed on 7 November 2002)

Flynn, Peter. (1998). *Understanding SGML and XML Tools: Practical Programs for Handling Structured Text.* Boston; Dordrecht; London: Kluwer, pp.212-235.

Follett, Brian. (1993). *Joint Funding Council's Libraries Review: Report (The Follett Report).* Bristol: HEFCE. Online. Available: http://www.ukoln.ac.uk/services/papers/follett/report/ (Accessed on 3 January 2004)

Folsom, Ed and Price, Kenneth M, editors. (2000). *The Walt Whitman Archive.* Online. Available: http://www.iath.virginia.edu/whitman/ (Accessed on 13 November 2003)

Fresco, Marc. (1996). *Long Term Preservation of Electronic Materials.* Report from JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib), organized by UKOLN, 27-28 November 1995, University of Warwick. British Library R&D Report, 6328. London: British Library. Online. Available: http://www.ukoln.ac.uk/services/papers/bl/rdr6238/paper.html (Accessed on 13 February 2004)

Freter, Todd. (1998a). *Beyond Text and Graphics: XML Makes Web Pages Function Like Applications.* Online. Available: http://www.sun.com/980414/xml/ (Accessed on 30 May 2001)

Freter, Todd. (1998b). *XML: Mastering Information on the Web.* Online. Available: http://www.sun.com/980310/xml/ (Accessed on 23 June 2001)

Frey, F and Reilly, J. (1999). *Digital Imaging for Photographic Collections: Foundations for Technical Standards*. Online. Available: http://www.rit.edu/~661www1/sub_pages/ digibook.pdf (Accessed on 21 November 2002)

Friedland, LeeEllen, editor. (1998). *American Memory DTD: for the Encoding of Full Texts of Historical Documents: SGML Tag Library. Online.* Available: http://memory.loc.gov/gc/lhbpr/03875/amtaglib.txt (Accessed on 8 August 2004)

Friedlander, Amy. (2002). The National Digital Information Infrastructure Preservation Program: Expectation, Realities, Choices and Progress to Date. *D-Lib Magazine*, 8(4). Online. Available: http://www.dlib.org/dlib/april02/friedlander/04friedlander.html (Accessed on 8 November 2002)

Fuhr, Norbert et al. (2001). Digital Libraries: a Gemroc Classification and Evaluation Scheme. In: Panos Constantopoulos and Ingeborg T. Sølvberg, editors, *Proceedings of the 5th European Conference: Research and Advanced Technology for Digital Libraries (ECDL 2001) Darmstardt, Germany, 4-9 September 2001.* Berlin; Heidelberg: Springer-Verlag, pp.187-199.

Gardner, Tracy. (2001). An Introduction to Web Services. *Ariadne,* 29. Online. Available: http://www.ariadne.ac.uk/issue29/gardner/intro.html (Accessed on 8 October 2001)

Garrett, John and Waters, Donald. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Online. Available: http://www.rlg.org/ArchTF/tfadi.index.htm (Accessed on 8 October 2001)

Gartner, Richard. (2002). *METS: Medata Encoding and Transmission Standard.* Techwatch Report, TSW 02-05. Online. Available: http://www.jisc.ac.uk/ index.cfm?name=techwatch_report_0205 (Accessed on 23 September 2003)

Getz, Ken et al. (1994). *Microsoft Access 2: Developer's Handbook.* San Francisco, California: Sybex, Chapter 2.

Gilliland-Swetland, Anne. (2000). *Introduction to Metadata: Setting the Stage.* Online. Available: http://www.getty.edu/research/conducting_research/standards/intrometadata/ 2_articles/index.html (Accessed on 4 December 2002)

GILS. (n.d.). Online. Available: http://www.gpoaccess.gov/index.html (Accessed on 2 August 2001)

Giordano, Richard. (1994). The Documentation of Electronic Texts Using Text Encoding Initiative Headers: an Introduction. *Library Resources and Technical Services,* 38:

389-402.

Givens, G. (2004). Universities' Fundraising Woes Reflect the Clash of Hope Against Reality. *The Economist* (17 April 2004): 26-29.

Goldfarb, Charles. (1990). *The SGML Handbook.* Oxford: Oxford University Press, pp. 5-8.

Goldfarb, Charles. (2004). *Biography: Dr. Charles F. Goldfarb.* Online. Available: http://www.sgmlsource.com/press/CGbioFull.htm (Accessed on 5 January 2004)

Goldfarb, Charles et al. (1997). *A Reader's Guide to the HyTime Standard.* Online. Available: http://www.hytime.org/papers/htguide.html (Accessed on 29 July 2005)

Google Inc. (2004). *Google Checks out Library Books.* Online. Available: http://www.google.co.uk/press/pressrel/print_library.html (Accessed on 11 March 2005)

Goossens, Michael and Saarela, Janne. (1995). *A Practical Introduction to SGML.* Online. Available: http://www.irb.hr/~cern/WWW/publications/sgmlen/sgmlen.html (Accessed on 24 April 2001)

Gorman, Michael. (1995). Five New Laws of Librarianship. *American Libraries*, September, pp. 784-785.

Gourley, Don. (2002). Integrating Systems with XML-based Web Services. In: Tennant, Roy, editor, *XML in Libraries*. New York: Neal-Schuman, pp.181-195.

Gourley, Don. (2003). *Library Portal Roles in a Shibboleth Federation.* Online. Available: http://shibboleth.internet2.edu/docs/gourley-shibboleth-library-portals-200310.html (Accessed on 3 December 2003)

Granger, Stewart. (1999). Metadata and Digital Preservation: a Plea for Cross-Interest Co-Operation. *VINE*, theme issue on Metadata, 117 (2): 24-29. Online. Available: http://dspace.dial.pipex.com/stewartg/metpres.html (Accessed on 7 October 2001)

Gredley, Ellen and Hopkinson, Alan. (1990). *Exchanging Bibliographic Data: MARC and Other International Formats.* London: Library Association, pp. 24-31 and 44-52.

Green, Brian. (2001). *Developments in Metadata: ONIX for Serials.* Paper presented in the Ejournals and the Web: Standards for Tomorrow. A NISO/BASIC program at ALA San Francisco, 17 June 2001. Online. Available: http://www.niso.org/news/releases/ PR-ejournal-web.html (Accessed on 18 November 2001)

Greenstein, Daniel and Thorin, Suzanne E. (2002). *The Digital Library: a Biography.* Washington, DC: Council on Library and Information Resources, Digital Library Federation. Online. Available: http://www.clir.org/pubs/reports/pub109/ contents.html (Accessed on 28 May 2003)

Griffin, Stephen M. (1999). Digital Libraries Initiative - Phase 2: Fiscal Year 1999 Awards. *D-Lib Magazine,* 5(7/8). Online. Available: http://www.dlib.org/dlib/july99/ 07griffin.html (Accessed on 23 December 2004)

Guittet. C, editor. (1985). *FORMEX: Formalized Exchange of Electronic Publications.* Brussels: Office for Official Publications of the European Community, New Technologies – Project Management Department. *passim.*

Gutheridge, Christopher. (2004). *EPrint Archives: What Technology is Involved? P*aper presented in the ePrint UK Bath Workshop, 6 February 2004.

Guthrie, Kevin M and Lougee, Wendy P. (1997). The JSTOR Solution Accessing and Preserving the Past. *Library Journal*, 122 (2): 42-44.

Haas, Hugo. (2003). *The Web Services Activity.* Online. Available: http://www.w3.org/2002/ws/#discussion (Accessed on 22 August 2003)

Habing, Thomas G et al. (2001). *Qualified Dublin Core using RDF for Sci-Tech Journal Articles.* Paper presented in the DC-2001 Conference, Tokyo, 22-26 October 2001. Online. Available: http://dli.grainger.uiuc.edu/publications/metadatacasestudy/ HabingDC2001.pdf (Accessed on 7 July 2003)

Haigh, Susan. (1998). *A Glossary of Digital Library Standards, Protocols and Formats.* Online. Available: http://www.nlc-bnc.ca/9/1/p1-253-e.html (Accessed on 23 July 2003)

Harmony. (n.d.). Online. Available: http://www.ilrt.bris.ac.uk/discovery/harmony/index.html (Accessed on 13 September 2002)

Harris, Jen. (1999). Perseus Project Poised to Receive $2.7 Million Grant. *The Observer,* 11 March. Online. Available: http://www.tufts.edu/as/stu-org/observer/ 1999/march11/news/2.htm (Accessed on 1 November 2002)

Heery, Rachel and Patel, Manjula. (2000). Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne*, 25. Online. Available: http://www.ariadne.ac.uk/issue25/app-profiles/ (Accessed on 8 July 2003)

Hey, Tony. (2004). Why Engage in e-Science? *Library + Information Update,* March. Online. Available:http://www.cilip.org.uk/publications/updatemagazine/archive /archive2004/march/update0403b.htm (Accessed on 21 February 2004)

Hockey, Susan. (1994). Electronic Texts in the Humanities: a Coming of Age. In: Brett Sutton, editor, *Literary Texts in an Electronic Age: Scholarly Implications and Library Services.* Papers presented in the Clinic on Library Applications of Data Processing, 10-12 April 1994. Urbana-Champaign, Illinois: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, pp.21-34.

Hockey, Susan. (2000). *Electronic Texts in the Humanities*. Oxford: Oxford University Press, Chapters 3, 5, 6 and 10.

Hockey, Susan. (2003). *An Interview with Professor Susan Hockey at SLAIS, UCL in January 2003.*

Hodge, Gail. (2001). *Metadata Made Simpler.* Bethesda, Maryland: National Information Standards Organization Press, p.3.

Hogan, Mike. (1997). *XML: the Foundation for the Future.* Online. Available: http://www.xml.org/xml/xml_foundation_future.shtml (Accessed on 2 March 2004)

Holden, Daniel. (1998). *JSTOR: 1999 and Beyond.* Online. Available: http://www.ariadne.ac.uk/issue18/jstor/ (Accessed on 21 February 2004)

Howlett, Scott and Dunmall, Jeff. (2000). *Beyond ASP: XML and XSL-Based Solutions Simplify Your Data Presentation Layer.* Online. Available: http://msdn.microsoft.com/msdnmag/issues/1100/BeyondASP/default.aspx (Accessed on 22 November 2002)

Iannella, Renato. (2001). Digital Rights Management (DRM) Architectures. *D-Lib Magazine,* 7(6). Online. Available: http://www.dlib.org/dlib/june01/iannella/06iannella.html (Accessed on 15 September 2001)

Iannella, Renato. (2002). *Open Digital Rights Language (ODRL), Version 1.1.* Online. Available: http://www.w3.org/TR/2002/NOTE-odrl-20020919/ (Accessed on 5 January 2004)

IFLA Study Group on the FRBR. (1998). *Functional Requirements for Bibliographic Records.* Final Report. Munich: K. G. Saur. Online. Available: http://www.ifla.org/VII/s13/frbr/frbr.htm (Accessed on 12 November 2004)

Information Infrastructure Task Force. (1994). *Libraries and the NII in Putting the Information Infrastructure to Work.* Committee on Applications and Technology Report, US Department of Commerce, May 1994. Online. Available: http://www.nist.gov/public_affairs/releases/g94-83.htm (Accessed on 10 September 2001)

INGENTA. (n.d.). Online. Available: http://www.ingenta.com/ (Accessed on 3 January 2004)

Intellor. (2001). *XML Adoption: Benefits and Challenges.* Online. Available: http://www.dad.be/library/pdf/intellor2.pdf (Accessed on 22 February 2004)

Internet2. (2001). Online. Available: http://www.internet2.edu/ (Accessed on 5 October 2001)

Internet2. (2005). Shibboleth Project. Online. Available: http://shibboleth.internet2.edu/ (Accessed on 14 April 2005)

ISO/TC46. (2004). *Resolutions ISO/TC46 Information and Documentation [2004-10-29].* Geneva: ISO, p.3.

ISO/TC46/SC4. (2004). *ISO/TC46/SC4 Approved resolutions [2004-10-27].* Geneva: ISO, p.2.

ITMT – Findings. (2001). *Internet Traffic Management Trial: Findings and Recommendations, University of Newcastle.* Online. Available: http://www.newcastle.edu.au/services/computing/iap/findings.pdf (Accessed on 9 October 2003)

Ivanov, Ivelin. (2003). *XQuery Implementation.* Online. Available: http://www.xml.com/pub/a/2003/10/01/xquery.html (Accessed on 3 January 2004)

James, Hamish. (2003). *Introduction to Creating Digital Resources.* Online. Available: http://www.ahds.ac.uk/visualarts/creating/information-papers/creating-introduction/index.htm (Accessed on 5 December 2003)

Jelliffe, Rick. (1998). *The XML & SGML Cookbook.* Upper Saddle River, New Jersey: Prentice

Hall PTR, p.19.

JISC. (1997). *JISC Circular 3/97 - Electronic Information Development Programme: eLib Phase 3.* London: JISC.

JISC. (2002). *Requirements for a Virtual Learning Environment.* Online. Available: http://www.jisc.ac.uk/index.cfm?name=mle_related_vle (Accessed on 5 January 2004)

John Wiley & Sons Ltd. (1999). *Reference Linking Service to Aid Scientists Conducting Online Research: Scientific and Scholarly Publishers Collaborate to Offer Ground-Breaking Initiative.* Online. Available: http://www.doi.org/ref-link-press-release-11-99.html (Accessed on 8 August 2005)

Johnston, Pete. (2002). *Collection-Level Description: Potential and Reality.* Collection Description Focus Briefing Day 2, British Library, St. Pancras, London, 14 May 2002. Online. Available: http://www.ukoln.ac.uk/cd-focus/presentations/realise/ (Accessed on 6 July 2003)

Johnston, Pete. (2005). What Are Your Terms? *Ariadne,* 43. Online. Available: http://www.ariadne.ac.uk/issue43/johnston/#top (Accessed on 10 July 2005)

Jørgensen, Paul H. (2000). *Z39.50, XML and RDF Applications.* ZIG Tutorial. Online. Available: http://www.loc.gov/z3950/agency/zig/meetings/texas/tutorials/xml-rdf.ppt (Accessed on 22 September 2003)

Jørgensen, Paul H. (2001). *XML Standards and Library Applications.* Paper presented in the ELAG 2001: Integrating Heterogeneous Resources, Prague, 6-8 June 2001. Online. Available: http://www.stk.cz/elag2001/Papers/Poul_HenrikJoergensen/Show.html (Accessed on 15 September 2003)

King, Philip and Woolls, David. (n.d.). *Creating and Using a Multilingual Parallel Concordancer.* Online. Available: http://xml.coverpages.org/kingCreatingConcordancer.html (Accessed on 9 October 2004)

Kling, Rob. (1999). What is Social Informatics and Why Does It Matter? *D-Lib Magazine,* 5(1). Online. Available: http://www.dlib.org/dlib/january99/kling/01kling.html (Accessed on 29 July 2003)

Koch, Traugott. (1999). *Adding Automatic Classification to a Robot-Generated Subject Index (DESIRE II) Demonstration.* Online. Available: http://www.lub.lu.se/tk/demos/korg9905-class.html (Accessed on 12 October 2001)

Lacher, Martin S and Decker, Stefan. (2001). *On the Integration of Topic Maps and RDF Data.* Paper presented in the 1st International Semantic Web Working Symposium (SWWS 01), Stanford University, Stanford, California, 29 July – 1 August 2001. Online. Available: http://www.semanticweb.org/SWWS/program/full/paper53.pdf (Accessed on 26 December 2003)

Lagoze, Carl and Payette, Sandra. (2000). Metadata: Principles, Practices, and Challenges. In:

Anne R. Kenney and Oya Y. Rieger, editors, *Moving Theory Into Practice: Digital Imaging for Libraries and Archives.* Mountain View, California: Research Libraries Group, pp.84-100.

Lamolinara, Guy. (1996). The Learning Page: Reaching K-12 Students On-line. *National Digital Library Program Periodical Report*, 9. Online. Available: http://lcweb.loc.gov/ndl/june-96.html (Accessed on 8 November 02 2002)

Lancaster, F W. (1978). *Toward Paperless Information Systems.* New York: Academic Press, Chapter 6.

Lander, Richard. (1997). *XML: the New Markup Wave.* Online. Available: http://xml.coverpages.org/landerXML.html (Accessed on 25 April 2001)

Lapeyre, Deborah and Usdin, Tommie. (1996). TEI and the American Memory Project at the Library of Congress. In *Proceedings of the 1st ACM International Conference on Digital Libraries, the Text Encoding Initiative Guidelines and Their Application to Building Digital Libraries, Bethesda, Maryland, 20-23 March 1996.* New York: ACM Press, pp. 184-185.

Lassila, Ora. (1997). *Introduction to RDF Metadata, W3C NOTE 13 November 1997.* Online. Available: http://www.w3.org/TR/NOTE-rdf-simple-intro-971113.html (Accessed on 22 October 2000)

Lassila, Ora and Swick, Ralph R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999.* Online. Available: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ (Accessed on 12 September 2000)

Lawrence Berkeley National Laboratory. (1997). *Workshop Report.* Joint Workshop on Metadata Registries. Online. Available: http://pueblo.lbl.gov/~olken/EPA/ Workshop/report.html (Accessed on 28 September 2003)

Le Boeuf, Patrick. (2003). *Brave New FRBR World.* Online. Available: http://www.ddb.de/news/pdf/papers_leboeuf.pdf (Accessed on 12 November 2004)

Lee, Edmund. (2005). Building Interoperability for United Kingdom Historic Environment Information Resources. *D-Lib Magazine*, 11(6). Online. Available: http://www.dlib.org/dlib/june05/lee/06lee.html (Accessed on 10 August 2005)

Lee, Stuart D. (2001). *Digital Imaging: a Practical Handbook.* London: Library Association Publishing, pp.64-66 and 103-109.

LeVan, Ralph. (1998). *Dublin Core and Z39.50.* Online. Available: http://es.dublincore.org/ documents/1998/02/02/dc-z3950/index.shtml (Accessed on 9 September 2003)

Levering, Mary. (1995). The Library and Copyright in the Digital Age. *National Digital Library Program Periodical Report*, 3. Online. Available: http://lcweb.loc.gov/ndl/oct-95.html (Accessed on 7 November 2002)

Library of Congress. (n.d.). *America, Russia and the Meeting of Frontiers*. Online. Available: http://international.loc.gov/intldl/mtfhtml/mfovrvw.html (Accessed on 9 January 2004)

Library of Congress. (1998a). *American Memory DTD for Historical Documents.* Online. Available: http://memory.loc.gov/ammem/techdocs/amdtd.html (Accessed on 17 June 2002)

Library of Congress. (1998b). *Challenges to Building an Effective Digital Library.* Online. Available: http://lcweb2.loc.gov/ammem/dli2/html/cbedl.html (Accessed on 7 October 2002)

Library of Congress. (1998c). *EAD Design Principles.* EAD Tag Library for Version 1.0. Online. Available: http://www.loc.gov/ead/tglib1998/tlprinc.html (Accessed on 22 January 2002)

Library of Congress. (1998d). *General Comments on Digital Reproductions of Textual Materials for American Memory Technical Notes: by Types of Material: Motion Pictures.* Online. Available: http://lcweb2.loc.gov/ammem/dli2/html/motion.html (Accessed on 20 October 2002)

Library of Congress. (1998e). *National Digital Library Program.* Online. Available: http://memory.loc.gov/ammem/dli2/html/lcndlp.html#Overview (Accessed on 16 July 2005)

Library of Congress. (1999). Library of Congress Kicks off Bicentennial Gifts to the Nation Project. *News from the Library of Congress, PR 99-049.* Online. Available: http://www.loc.gov/today/pr/1999/99-049.html (Accessed on 8 January 2004)

Library of Congress. (2001a). *METS: Metadata Encoding and Transmission Standard.* Online. Available: http://www.loc.gov/standards/mets/ (Accessed on 28 May 2001)

Library of Congress. (2001b). *Spain, the United States & the American Frontiers: Historias Paralelas*. Online. Available: http://international.loc.gov/intldl/eshtml/ (Accessed on 9 January 2004)

Library of Congress. (2002). The Library of Congress: Report to the Digital Library Federation. *Digital Library Federation Newsletter*, 3(1). Online. Available: http://www.diglib.org/pubs/news03_01/lc.htm (Accessed on 9 March 2003)

Library of Congress. (2003a). *Development of the Encoded Archival Description DTD.* Online. Available: http://www.loc.gov/ead/eaddev.html (Accessed on 9 January 2004)

Library of Congress. (2003b). *The Library of Congress/Ameritech National Digital Library Competition (1996-1999).* Online. Available: http://memory.loc.gov/ammem/award/index.html (Accessed on 9 January 2004)

Library of Congress. (2003c). *METS: an Overview and Tutorial.* Online. Available: http://www.loc.gov/standards/mets/METSOverview.v2.html (Accessed on 31 January 2004)

Library of Congress. (2004). *Election 2002 Web Archive Cataloging & Description.* Online. Available: http://www.loc.gov/minerva/collect/elec2002/catalog.html (Accessed on 3 September 2004)

Library of Congress. (2005a). *METS Implementation Registry.* Online. Available: http://sunsite.berkeley.edu/mets/registry/ (Accessed on 17 August 2005)

Library of Congress. (2005b). *Standards at the Library of Congress.* Online. Available: http://www.loc.gov/standards/ (Accessed on 11 August 2005)

Library of Congress Copyright Office. (2003). *Copyright Office Electronic Registration, Recordation, and Deposit System.* Online. Available: http://www.copyright.gov/cords/ (Accessed on 28 October 2003)

Library of Congress Z39.50 Maintenance Agency. (2002). *Z39.50 Holdings Schema.* Online. Available: http://lcweb.loc.gov/z3950/agency/defns/holdings.html (Accessed on 26 October 2003)

Light, Richard. (1996). *Project CHIO Tagging Guid, Version 1.0.* Online. Available: http://www.cimi.org/public_docs/tagging_guide/tg.htm (Accessed on 22 January 2002)

Linden, Greger and Heinonen, Oskari. (1997). *DocMan: Document Management Research Group.* Online. Available: http://www.cs.helsinki.fi/research/rati/docman.html#PU (Accessed on 20 October 2004)

Linden, Greger. (2004). *DocMan.* E-mail to Naicheng Chang. 21 October 2004.

Lougee, Wendy P. (1998). The University of Michigan Digital Library Program: a Retrospective on Collaboration within the Academy. *Library Hi Tech*, 16(1): 51-59.

Lozano, Fernando. (n.d.). *EDM/2: Introduction to Relational Database Design.* Online. Available: http://www.edm2.com/0612/msql7.html (Accessed on 12 June 2001)

Lynch, C, editor. (1998). *A White Paper on Authentication and Access Management Issues in Cross-organizational Use of Networked Information Resources.* Coalition for Networked Information. Revised Discussion Draft of 14 April 1998. Online. Available: http://www.cni.org/projects/authentication/authentication-wp.html (Accessed on 15 December 2002)

M25 Consortium. (n.d.). *The M25 Consortium of Academic Libraries.* Online. Available. http://www.m25lib.ac.uk/ (Accessed on 6 January 2004)

MacColl, John. (2002). Electronic Theses and Dissertations: a Strategy for the UK. *Ariadne*, 32. Online. Available: http://www.ariadne.ac.uk/issue32/theses-dissertations (Accessed on 17 November 2003)

Maler, Eve. (n.d.). *XML Linking: State of the Art.* Online. Available: http://wwws.sun.com/software/xml/developers/xlink/ (Accessed on 7 November 2002)

Malhotra, Ashok et al., editors. (2003). *XML Syntax for XQuery 1.0 (XqueryX), W3C Working Draft 7 June 2001.* Online. Available: http://www.w3.org/TR/xqueryx (Accessed on 29

December 2003)

MALVINE. (2003). Online. Available: http://www.malvine.org/ (Accessed on 12 August 2005)

Manola, Frank, editor. (1997). *Object Model Features Matrix, Technical Committee H7, Doc. No.: X3H7-93-007v12b*. Online. Available: http://www.objs.com/x3h7/fmindex.htm (Accessed on 28 June 2003)

Marchionini, Gary. (2000). Evaluating Digital Libraries: a Longitudinal and Multifaceted View. *Library Trends*, 49(2): 304-333.

Marchionini, Gary and Crane, H. (1994). Evaluating Hypermedia and Learning: Methods and Results From the Perseus Project. *ACM Transactions on Information Systems*, 12(1): 5-34.

Marchionini, Gary and Maurer, Hermann. (1995). The Role of Digital Libraries in Teaching and Learning. *Communications of the ACM*, 38(4): 67-75.

Marchionini, Gary et al. (1998). Interfaces and Tools for the Library of Congress National Digital Library Program. *Information Processing and Management*, 5(34): 535-555.

Marchionini, Gary et al. (2000). *Final Evaluation Report on the Perseus Project Publication Model 1997-2000*. Paper submitted to the Fund for the Improvement of Post Secondary Education, 15 August 2000. Online. Available: http://www.ils.unc.edu/~march/perseus/final_report.pdf (Accessed on 30 October 2002)

Martínez, José et al. (2002). MPEG-7: the Generic Multimedia Content Description Standard. (part 1) *IEEE Multimedia*, 9(2): 78-87.

MASTER. (2001). Online. Available: http://www.cta.dmu.ac.uk/projects/master/index.html (Accessed on 14 November 2001)

MATRIX. (2003). Online. Available: http://matrix.msu.edu/ (Accessed on 3 January 2004)

McDonough, Jerome. (2003). *METS: a Status Report*. CNI Task Force Meeting, Washington, DC, 28-29 April 2003. Online. Available: http://www.cni.org/tfms/2003a.spring/powerpoints/PPT-METS-McDonough.ppt (Accessed on 23 September 2003)

McDonough, Jerome and Proffitt, Merrilee. (2001). *METS: Metadata Encoding for Digital Objects*. Online. Available: http://www.cni.org/tfms/2001b.fall/handout/METS-JMcDonough2001Ftf.pdf (Accessed on 23 September 2003)

McHugh, Jason et al. (1997). Lore: a Database Management System for Semi-Structured Data. *SIGMOD Record*, 26(3): 54-66.

McHugh, Jason and Widom, Jennifer. (1999). Query Optimization for XML. In: Malcolm P. Atkinson et al., editors, *Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, 7-10 September, 1999*. San Francisco, California: Morgan Kaufmann, pp.315-326.

MEDLANE. (2002). Online. Available: http://laneweb.stanford.edu:2380/wiki/medlane/

overview (Accessed on 2 August 2002)

Mertz, David. (2001). *XML Matters: Comparing W3C XML Schemas and Document Type Definitions (DTDs).* Online. Available: http://www-106.ibm.com/developerworks/library/x-matters7.html (Accessed on 28 May 2001)

Miller, Dick. (2002). *Adding Luster to Librarianship: XML as an Enabling Technology.* Online. Available: http://elane.stanford.edu/laneauth/Luster.html (Accessed on 14 October 2004)

Miller, Dick R and Clarke, Kevin S. (2003). *Putting XML to Work in the Library: Tools for Improving Access and Management.* Chicago, Illinois: American Library Association, 205pp.

Miller, Eric. (1998). An Introduction to Resource Description Framework. *D-Lib Magazine*, May. Online. Available: http://www.dlib.org/dlib/may98/miller/05miller.html (Accessed on 19 December 2000)

Millman, David. (1999). Cross-Organizational Access Management: a Digital Library Authentication and Authorization Architecture. *D-Lib Magazine*, 5(11). Online. Available: http://www.dlib.org/dlib/march97/sesame/03clews.html (Accessed on 28 September 2003)

Mimno, David. (2004). *Requesting Update Information on PDL XML Activities.* E-mail to Naicheng Chang. 8 September 2004.

MLA. (2005). *The People's Network.* Online. Available: http://www.peoplesnetwork.gov.uk/about.html (Accessed on 14 August 2005)

MOA2. (1998). *The Making of American II Testbed Project White Paper, Version 2.* Online. Available: http://sunsite.berkeley.edu/moa2/wp-v2.html (Accessed on 24 December 2002)

MODELS. (1999). Online. Available: http://www.ukoln.ac.uk/dlis/models/clumps/ (Accessed on 22 September 2003)

MODS. (2002). Online. Available: http://www.loc.gov/standards/mods/ (Accessed on 22 July 2002)

Morgan, R L et al. (2004). Federated Security: the Shibboleth Approach. *Educause Quarterly*, 27(4). Online. Available: http://www.educause.edu/apps/eq/eqm04/eqm0442.asp?bhcp=1 (Accessed on 15 April 2005)

Morrison, Alan et al. (2000). *(Oxford Text Archive) Creating and Documenting Electronic Texts.* Oxford: Oxbow Books, pp.27-49.

MPEG. (n.d.). Online. Available: http://www.chiariglione.org/mpeg/ (Accessed on 17 February 2004)

MPEG-7 DDL. (n.d.). *Overview of the MPEG-7 Standard ISO/IEC JTC1/SC29/WG11 N4031.* Online. Available: http://vision.hanyang.ac.kr/main/Mpeg7/overview/M7_ddl.html (Accessed on 11 November 2001)

Muench, Steve. (n.d.). *Using XML and Relational Databases for Internet Applications*. Online. Available: http://otn.oracle.com/tech/xml/htdocs/relational/paper.html (Accessed on 22 June 2001)

Muller, Charles, editor. (2001). *Digital Dictionary of Buddhism*. Online. Available: http://www.acmuller.net/ddb/ (Accessed on 11 November 2001)

Muller, Charles and Beddow, Michael. (2002). Moving into XML Functionality: the Combined Digital Dictionaries of Buddhism and East Asian Literary Terms. *Journal of Digital Information,* 3 (2). Online. Available: http://jodi.ecs.soton.ac.uk/Articles/ v03/i02/Muller/#2 (Accessed on 21 May 2005)

Musciano, Chuck and Kennedy, Bill. (1997). *HTML: the Definitive Guide*. 2nd edition, Cambridge; Sebastopol: O'Reilly, Chapter 1.

Myrick, Leslie. (2002). Harnessing Oracle and XT for Finding Aid Dissemination and Search. In: Tennant, Roy, editor, *XML in Libraries*. New York: Neal-Schuman, pp.45-57.

National Academy of Sciences. (2000). *C21: a Digital Strategy for the Library of Congress*. Washington, DC: National Academy Press, pp.90-104, 105-121 and 204-210.

National Library of New Zealand. (2001). Facing the Challenge of Digital nformation.*CDNLAO Newsletter*, 42. Online. Available: http://www.ndl.go.jp/en/publication/cdnlao/ 042/424.html (Accessed on 17 December 2002)

National Science Council. (2004). *Consortium on Core Electronic Resources in Taiwan (CONCERT)*. Online. Available: http://www.stic.gov.tw/fdb/consortium/index.html (Accessed on 3 February 2004)

National Science Digital Library. (2003). *Sharing your Stories: Sustaining Collaborations*. NSDL Annual Meeting. Online. Available: http://swiki.dlese.org/nsdl2003/1 (Accessed on 13 June 2004)

Naughton, John. (2000). *A Brief History of the Future: the Origins of the Internet*. London: Phoenix, p.240.

NDMSO. (2001). *MARC to Dublin Core Crosswalk*. Online. Available: http://www.loc.gov/ marc/marc2dc.html (Accessed on 22 September 2003)

NDMSO. (2003). *MARC Standards*. Online. Available: http://www.loc.gov/marc/ (Accessed on 8 January 2004)

NDMSO. (2005). *MARCXML: MARC 21 XML Schema*. Online. Available: http://www.loc.gov/ standards/marcxml/ (Accessed on 8 August 2005)

Nelson, Theodor H. (1974). *Dream Machines: New Freedoms through Computer Screens: a Minority Report*. [Chicago, Illinois]: The Author, 1974. passim.

NetLibrary. (2005). Online. Available: http://legacy.netlibrary.com/about_us/company_info/ (Accessed on 3 February 2005)

Network Inference. (2005). Online. Available: http://www.networkinference.com/ (Accessed on

12 August 2005)

New, Juliet. (2000). The World's Greatest Dictionary Goes Online. *Ariadne,* 23. Online. Available: http://www.ariadne.ac.uk/issue23/oed-online/ (Accessed on 5 May 2002)

NINCH. (2002). *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials.* Washington, DC: NINCH. Online. Available:http://www.nyu.edu/its/humanities/ninchguide/index.html (Accessed on 23 November 2004)

NISO. (1999). *NISO/CLIR/RLG Technical Metadata Elements for Images Workshop.* Washington, DC, 18-19 April 1999. Online. Available: http://www.niso.org/news/events_workshops/image.html (Accessed on 11 October 2001)

NISO. (2002). *NISO Circulation Interchange Protocol (NCIP).* Online. Available: http://www.niso.org/standards/resources/z3983pt1rev1.pdf (Accessed on 20 October 2004)

Nixon, William. (2003). DAEDALUS: Initial Experiences with EPrints and DSpace at the University of Glasgow. *Ariadne*, 37. Online. Available: http://www.ariadne.ac.uk/issue37/nixon (Accessed on 17 November 2003)

NLM. (2003). *Journal Publishing DTD.* Online. Available: http://dtd.nlm.nih.gov/publishing/ (Accessed on 25 February 2004)

OAI. (n.d.). Online. Available: http://www.openarchives.org (Accessed on 3 September 2003)

OCLC. (1995). *History of the Dublin Core Metadata Initiative.* Online. Available: http://dublincore.org/about/history/ (Accessed on 21 February 2004)

OCLC. (2002). *A Metadata Framework to Support the Preservation of Digital Objects.* Report by the OCLC/RLG Working Group on Preservation Metadata, June 2002. Online. Available: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf (Accessed on 18 November 2002)

Ogbuji, Uche. (2002). *Using RDF with SOAP: beyond Remote Procedure Calls.* Online. Available: http://www-106.ibm.com/developerworks/webservices/library/ws-soaprdf/ (Accessed on 8 March 2004)

ONIX. (n.d.). Online. Available: http://www.editeur.org/onix.html (Accessed on 18 November 2001)

OSDLS. (1999). *Open Source Digital Library System.* Online. Available: http://www.library.arizona.edu/users/jfrumkin/osdls/welcome.html (Accessed on 22 September 2004)

OSS4LIB. (2005). *Open Source Systems for Libraries.* Online. Available: http://www.oss4lib.org/ (Accessed on 12 August 2005)

Ossenbruggen, Jacco van et al. (2001). Towards Second and Third Generation Web-Based Multimedia. In: *The 10th International World Wide Web Conference on World Wide Web (WWW10), Hong Kong, 1-5 May 2001.* New York: ACM Press, pp.479-488.

OzAuthors. (n.d.). Online. Available: http://www.ozauthors.com.au/ (Accessed on 4 December 2003)

Paepcke, Andreas et al. (1998). Interoperability for Digital Libraries Worldwide. *Communications of the ACM,* 41(4): 33-43.

Palowitch, Casey and Stewart, Darin. (1995). *Automating the Structural Markup Process in the Conversion of Print Documents to Electronic Text.* Online. Available: http://www.csdl.tamu.edu/DL95/papers/palowitc/palowitc.html (Accessed on 13 September 2003)

Pawson, Dave. (2001). *XSLT Questions and Answers*. Online. Available: http://www.dpawson.co.uk/xsl/sect2/sect21.html (Accessed on 19 June 2001)

PDL. (n.d.). *Perseus Digital Library.* Online. Available: http://www.perseus.tufts.edu/ (Accessed on 16 July 2005)

Penn Database Research Group. (2001). *Semistructured Data and XML*. Online. Available: http://db.cis.upenn.edu/Research/SS_XML.html (Accessed on 25 April 2001)

PERSEO. (n.d.). Online. Available: http://www.disam.upm.es/vision/projects/perseo/index.html (Accessed on 16 July 2003)

Piattini, Mario and Diaz, Oscar, editors. (2000). *Advanced Database Technology and Design.* Boston, Massachusetts; London: Artech House, pp.3-8.

Pinfield, Stephen. (2001). Managing Electronic Library Services: Current Issues in UK Higher Education Institutions. *Ariadne*, 29. Online. Available: http://www.ariadne.ac.uk/ issue29/pinfield/intro.html (Accessed on 28 Janurary 2003)

Powell, C K. (2004). *Requesting Update Information on UMDL XML Activities.* E-mail to Naicheng Chang. 17 August 2004.

Powell, C K and Kerr, Nigel. (1997). SGML Creation and Delivery: the Humanities Text Initiative. *D-Lib Magazine*, July/August. Online. Available: http://www.dlib.org/dlib/ july97/humanities/07powell.html (Accessed on 9 June 2001)

Price-Wilkin, John. (1999). *Moving the Digital Library from "Project" to "Production".* Paper presented in the DLW'99, Tsukuba, Japan, February 1999. Online. Available: http://jpw.umdl.umich.edu/pubs/japan-1999.htm (Accessed on 27 October 2002)

Price-Wilkin, John. (2002). *Broader Rather Than Deeper: TEI and the Importance of Relevance to the Digital Library Community.* Paper presented in the 2nd Annual Meeting of the TEI Consortium, Chicago, Illinois, 11-12 October 2002. Online. Available: http://www.tei-c.org.uk/Publicity/chicago.html (Accessed on 28 November 2002)

Prud'hommeaux, Eric. (2001). *RDF Model for WSDL.* Online. Available:

http://www.w3.org/2002/02/21-WSDL-RDF-mapping/ (Accessed on 8 March 2004)

Public Library of Science. (n.d.). *PLOS Publishing Model.* Online. Available: http://www.plos.org/journals/model.html (Accessed on 2 February 2004)

Quint, Barbara. (2002). *BioMed Central Begins Charging Authors and Their Institutions for Article Publsihing.* Online. Available: http://www.infotoday.com/newsbreaks/nb020107-1.htm (Accessed on 2 February 2004)

Ragnarsdottir, Sigrun. (1999). *Data Storage and Delivery on the Web and the Role of XML.* Online. Available: http://www.cs.cornell.edu/cs632-sp99/surveys/sigrun.doc (Accessed on 9 January 2001)

Rahtz, Sebastian. (2004). *The P5 Version of the TEI.* Online. Available: http://www.tei-c.org.uk/P5/p5.html (Accessed at 23 February 2005)

Ranganathan, Shiyali Ramamrita. (1931). *The Five Laws of Library Science.* Madras: Madras Library Association, 464pp.

Rauber, Andreas and Tjoa, A Min. (2001). *User Interfaces for Digital Libraries.* Online. Available: http://www.faw.uni-linz.ac.at/online/papers/2001ocg/rau_ocg01.pdf (Accessed on 6 July 2003)

Ray, Louise. (1994). Evaluating Library Automation Systems for Archive Management. In: *Proceedings of the 18th International On-Line Information Meeting, London. 6-8 December, 1994.* Oxford: Learned Information, pp.499-504.

Ream, Dan. (1993). The University of Virginia's Electronic Text Center: an Interview with David Seaman. *Virginia Librarian,* 39(2). Online. Available: http://etext.lib.virginia.edu/articles/VirgLib/virglib.html (Accessed on 12 October 2004)

Renear, Allen H. (1997). The Digital Library Research Agenda: What's Missing -- and How Humanities Textbase Projects Can Help. *D-Lib Magazine,* July/August. Online. Available: http://www.dlib.org/dlib/july97/07renear.html (Accessed on 6 July 2003)

Renear, Allen H et al. (2002). Towards a Semantics for XML Markup. In: R. Furuta, J. I. Maletic, and E. Munson, editors, *Proceedings of the 2002 ACM Symposium on Document Engineering, McLean, Virginia, 8-9 November 2002.* New York: ACM Press, pp.119-126.

Renear, Allen H et al. (2003). XML Semantics and Digital Libraries. In: Catherine C. Marshall, Geneva Henry, and Lois Delcambre, editors, *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, 27-31 May 2003.* New York: ACM Press, pp. 303-305.

Re:source. (2005). Online. Available: http://www.resource.gov.uk/ (Accessed on 12 August 2005)

Rhyno, Art. (2002a). Building XML Databases with Zope and Castor. In: Tennant, Roy, editor, *XML in Libraries.* New York: Neal-Schuman, pp.117-131.

Rhyno, Art. (2002b). XML and Relational Databases: Uses and Opportunities for Libraries. *OCLC Systems & Services,* 18(2): 97-103.

Rick, Beaubien. (n.d.). *METS Overview Provisions.* Online. Available: http://www.loc.gov/standards/mets/presentations.html (Accessed on 14 December 2002)

RLG. (1998). *RLG Working Group on Preservation Issues of Metadata: Final Report.* Mountain View, California: Research Libraries Group. Online. Available: http://www.rlg.org/preserv/presmeta.html (Accessed on 8 October 2001)

Robie, Jonathan. (1999). *XML Query Language (XQL).* Online. Available: http://www.ibiblio.org/xql/xql-proposal.html (Accessed on 18 June 2001)

Robinson, Peter. (1993). *The Digitization of Primary Textual Sources.* Oxford: Office for Humanities Communication, pp. 4-5.

Robinson, Peter and Solopova, Elizabeth. (n.d.). *The Canterbury Tales Project.* Online. Available: http://www.ucalgary.ca/~scriptor/chaucer/rob.html (Accessed on 14 October 2004)

Robinson, Peter et al. (1999). *Initiatives Towards a Standard Encoding for Manuscript Descriptions.* Paper presented in the Conference of Digital Resources for the Humanities (DRH '99), King's College London, 12-15 September 1999. Online. Available: http://www.kcl.ac.uk/humanities/cch/drhahc/drh/abst46.htm (Accessed on 14 November 2001)

Rollitt, Karen et al. (2002). Using Dublin Core for DISCOVER: a New Zealand Visual Art and Music Resource for Schools. In: *Proceedings of International Conference on Dublin Core and Metadata for e-Communities: Supporting Diversity and Convergenc. Florence, Italy, 13 -17 October 2002.* Florence: Firenze University Press, pp. 251-255.

Roman de la Rose Project. (n.d.). *Text Encoding for the Roman de la Rose.* Project of the Milton S. Eisenhower Library of the Johns Hopkins University and the Pierpont Morgan Library. Online. Available: http://rose.mse.jhu.edu/pages/tagging.htm#top (Accessed on 9 December 2002)

Romary, Laurent, et al. (n.d.). *The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purpose.* Online. Available: http://xml.coverpages.org/lpcp.html (Accessed on 9 October 2004)

Rothenberg, Jeff. (1999). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Online. Available: http://www.clir.org/pubs/reports/rothenberg/contents.html (Accessed on 8 October 2001)

Rowley, Jennifer. (1998). *The Electronic Library.* London: Library Association Publishing, pp.19-20.

Rusbridge, Chris. (1998). Towards the Hybrid Library. *D-Lib Magazine*, July/August. Online. Available: http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html (Accessed on 6 January 2004)

Sairamesh, J et al. (1996). Economic Framework for Pricing and Charging in Digital Libraries. *D-Lib Magazine*, February. Online. Available: http://www.dlib.org/dlib/february96/ forth/02sairamesh.html (Accessed on 12 October 2001)

Sall, Ken. (1998). *XML: Structuring Data for the Web: an Introduction.* Online. Available: http://www.wdvl.com/Authoring/Languages/XML/Intro/ (Accessed on 24 January 2001)

Sankar, James and Garibyan, Mariam. (2005). Shibboleth Installation Workshop. *Ariadne*, 42. Online. Available: http://www.ariadne.ac.uk/issue42/shibboleth-rpt/ (Accessed on 15 April 2005)

Schauble, Peter and Smeaton, Alan F, editors. (1998). *An International Research Agenda for Digital Libraries.* Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research, 12 October 1998. Online. Available: http://www.iei.pi.cnr.it/DELOS/REPORTS/Brussrep.htm (Accessed on 26 July 2001)

Schmidt, Heidi et al. (2002). Building Digital Tobacco Industry Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D-Lib Magazine*, 8(9). Online. Available: http://www.dlib.org/ dlib/september02/ schmidt/09schmidt.html (Accessed on 10 February 2003)

School of Information and Library Studies, University of Michigan. (1991). *Information and People: a Campus Dialogue on the Challenges of Electronic Information: Committee Reports Prepared in Support of the Information Symposium, March 1991.* Ann Arbor, Michigan: University of Michigan, School of Information and Library Studies, 1991, 1 volume (no pagination).

SCONUL. (2005). Online. Available: http://www.sconul.ac.uk/pubs_stats/statques.html (Accessed on 13 August 2005)

Seaman, David. (1994). *Campus Publishing in Standardized Electronic Formats -- HTML and TEI.* Online. Available: http://etext.lib.virginia.edu/articles/arl/dms-arl94.html (Accessed on 13 October 2004)

Searle, Steven J. (2002). *A Brief History of Character Codes in North America,Europe, and East Asia.* Online. Available: http://tronweb.super-nova.co.jp/characcodehist.html (Accessed on 27 January 2004)

Sexton, Anna. (2003). *LEADERS, Progress Report, 4.* Online. Available: http://www.ucl.ac.uk/ leaders-project/Papers/progress4.htm (Accessed on 3 September 2003)

SGML Users' Group. (1990). A Brief History of the Development of SGML. *SGML NEWSWIRE,* 11 June. Online. Available: http://xml.coverpages.org/sgmlhist0.html (Accessed on 11 October 2004)

Shafer, Keith. (1998). Mantis Projects Provides a Toolkit for Cataloging. *OCLC Newsletter*, 236. Online. Available: http://digitalarchive.oclc.org/da/ViewObject.jsp?fileid=0000001718:000000044068&reqid=3456 (Accessed on 23 January 2002)

Shelley, E P and Johnson, B D. (1995). Metadata: Concepts and Models. In: *Proceedings of the 3rd National Conference on the Management of Geoscience Information and Data, Adelaide, Australia, 18-20 July 1995.* Glenside, South Australia: Australian Mineral Foundation. pp.4.1-4.5.

Shepherd, Peter T. (2004). COUNTER: Towards Reliable Vendor Usage Statistics. *VINE*, 34 (4): 184-189.

Sigel, Alexander. (2000). *Towards Knowledge Organization with Topic Maps*. Paper presented in the XML EUROPE 2000, Paris, 12-16 June 2000. Online. Available: http://www.gca.org/papers/xmleurope2000/papers/s22-02.html (Accessed on 15 October 2001)

Silberschatz, Abraham et al. (1997). *Database System Concepts*. 3rd edition, New York; London: McGraw-Hill, Chapter 9.

SMIL. (2001). Online. Available: http://www.w3.org/TR/2001/REC-smil20-20010807/ (Accessed on 11 November 2001)

Smiraglia, Richard P, editor. (1990). *Describing Archival Materials: the Use of the MARC AMC Format.* New York: Haworth Press, 228pp.

Smith, David A. (2001). *Linking and Gathering: Automatic Hypertext in the Perseus Digital Library.* Paper presented in the ACH/ALLC Conference, New York, New York University, 13-16 June 2001. Online. Available: http://www.nyu.edu/its/humanities/ach_allc2001/papers/smith-david/ (Accessed on 21 May 2005)

Smith, David A et al. (2000). Management of XML Documents in an Integrated Digital Library. *Markup Languages: Theory and Practice*, 2 (3): 205-214.

Snowhill, Lucia. (2001). E-Books and Their Future in Academic Libraries. *D-Lib Magazine*, 7 (7/8). Online. Available: http://www.dlib.org/dlib/july01/snowhill/07snowhill.html (Accessed on 1 February 2004)

SOAP. (2003). Online. Available: http://www.w3.org/TR/SOAP/ (Accessed on 10 September 2003)

Sperberg-McQueen, C M. (1998). *XML and What It Will Mean for Libraries.* Online. Available: http://tigger.uic.edu/~cmsmcq/talks/teidlf1.html (Accessed on 8 January 2002)

Sperberg-McQueen, C M and Burnard, Lou, editors. (1999). *The Guidelines for Electronic Text Encoding and Interchange (TEI P3).* Chapters 2 and 5. Online. Available: http://www.hti.umich.edu/t/tei/ (Accessed on 18 January 2002)

Sperberg-McQueen, C M and Burnard, Lou, editors. (2002). *The Guidelines for Electronic Text*

Encoding and Interchange (TEI P4). Online. Available: http://www.tei-c.org/P4X/ (Accessed on 18 November 2002)

Staken, Kimbro. (2001). *Introduction to Native XML Databases.* Online. Available: http://www.xml.com/pub/a/2001/10/31/nativexmldb.html (Accessed on 13 July 2003)

Stanford Digital Library Project. (1997). *Quarterly Report Stanford Digital Library Project.* Online. Available: http://www-diglib.stanford.edu/diglib/pub/reports/annuals/aug97.html (Accessed on 3 September 2003)

Staples, Thornton and Wayland, Ross. (2000). Virginia Dons FEDORA: a Prototype for a Digital Object Repository. *D-Lib Magazine*, 6(7/8). Online. Available: http://www.dlib.org/dlib/july00/staples/07staples.html (Accessed on 15 August 2001)

Star, Susan Leigh and Bishop, Ann. (1996). Social Informatics of Digital Library Use and Infrastructure. In: Martha E. Williams, editor, *Annual Review of Information Science and Technology*. White Plains, New York: Knowledge Industry Publications, pp. 301-401.

Sun Microsystems. (2001). *Using SQL3 Datatypes.* The Java Tutorial in New Features in the JDBC 2.0 API. Online. Available: http://java.sun.com/docs/books/tutorial/jdbc/jdbc2dot0/sql3.html (Accessed on 28 June 2001)

SWWS. (2001). Online. Available: http://www.semanticweb.org/SWWS/ (Accessed on 28 December 2003)

Taylor, R S. (1975). Pattern Toward a User-Centered Academic Library. In: E.J. Josey, editor, *New Dimensions for Academic Library Service.* Metuchen, New Jersey: Scarecrow Press, p.299.

TEI. (2001). *New Left Review*. Online. Available: http://www.tei-c.org.uk/Applications/apps-nl01.html (Accessed on 3 July 2001)

TEL. (2003). *The European Library (TEL): the Gate to Europe's knowledge.* Online. Available: http://www.europeanlibrary.org/ (Accessed on 23 February 2003)

Tennant, Roy. (2002a). Publishing Books Online at eScholarship. In: Tennant, Roy, editor, *XML in Libraries*. New York: Neal-Schuman, pp.101-114.

Tennant, Roy, editor. (2002b). *XML in Libraries*. New York: Neal-Schuman, 213pp.

TopicMaps. Org. (2001). *TopicMaps.Org Specification, Version 1.0*. Online. Available: http://www.topicmaps.org/xtm/1.0/ (Accessed on 15 October 2001)

TREC. (2003). Online. Available: http://trec.nist.gov/ (Accessed at 23 March 2003)

Tuck, William R. (1996). *Document Encoding Formats for On-demand Publishing.* LITC Report No.5. London: Library Information Technology Centre, South Bank University, pp.21-27.

Tudhope, Douglas and Cunliffe, Daniel. (1999). Semantically Indexed Hypermedia: Linking Information Disciplines. *ACM Computing Surveys*, 31(4es).

Turner, Chris. (2003). *Developing a Generic Toolkit: Architecture and Technology Issues.* Paper presented in the ACH/ ALLC Conference, Athens, the University of Georgia, 29 May - 2 June 2003. Online. Available: http://www.ucl.ac.uk/leaders-project/Papers/ALLCCT.ppt (Accessed on 17 October 2003)

Turner, Linda. (1990). A Brief History of the Development of SGML. *SGML NEWSWIRE*, 11 June. Online. Available: http://www.oasis-open.org/cover/sgmlhist0.html (Accessed on 10 September 2002)

UIUC. (2002a). *Digital Library Components.* Online. Available:http://www.cic.uiuc.edu/ groups/LibraryInfoTechDirectors/archive/Report/ DigitalLibraryComponents0419.pdf (Accessed on 15 July 2003)

UIUC. (2002b). *University of Illinois at Urbana-Champaign D-Lib Test Suite Project (1999-2001): Final Report.* Online. Available: http://dli.grainger.uiuc.edu/idli/ progress_reports/final_report.htm (Accessed on 1 December 2003)

Ullman, Jeffrey D. (1988). *Principles of Database and Knowledge-Base System.* Rockville, Maryland: Computer Science Press, Volume1, pp.74-77.

UMDL. (2003). *UM Digital Library Production Service: UMDL Texts.* Online. Available: http://www.hti.umich.edu/cgi/t/text/text-idx?tpl=home.tpl (Accessed on 10 August 2005)

Unicode, Inc. (2005). *What Is Unicode?* Online. Available: http://www.unicode.org/ standard/WhatIsUnicode.html (Accessed on 30 July 2005)

University of Michigan. (n.d.). *Library Information Technology.* Online. Available: http://www.lib.umich.edu/lit/ (Accessed on 16 July 2005)

University of Waikato. (2005). Online. Available: http://www.waikato.ac.nz/library/ (Accessed on 12 August 2005)

van der Vlist, Eric. (2000). *XML Linking Technologies.* Online. Available: http://www.xml.com/ pub/a/2000/10/04/linking/index.html (Accessed on 27 December 2003)

VRML Consortium Inc. (1997). *The Virtual Reality Modeling Language.* Online. Available: http://www.web3d.org/x3d/specifications/vrml/vrml97/ (Accessed on 7 August 2005)

W3C DOM Working Group. (2005). *W3C Document Object Model.* Online. Available: http://www.w3c.org/DOM/#what (Accessed on 12 August 2005)

W3C HTML Working Group. (2002). *XHTML 1.0: the Extensible HyperText Markup Language (Second Edition).* Online. Available: http://www.w3.org/TR/xhtml1/ (Accessed on 10 December 2003)

W3C RDF. (2004). Online. Available: http://www.w3.org/RDF/ (Accessed on 4 December 2004)

W3C SVG Working Group. (2005). *Scalable Vector Graphics: XML Graphics for the Web.* Online. Available: http://www.w3.org/Graphics/SVG/ (Accessed on 7 August 2005)

W3C XML Core Working Group. (1997). *XML Representation of a Relational Database.*

Online. Available: http://www.w3c.org/XML/RDB.html (Accessed on 31 July 2001)

Warburton, Yvonne L. (2004). *OED Online Technology.* E-mail to Naicheng Chang. 22 October 2004.

Watermarking World. (2005). *Digital Watermarking Frequently Asked Questions (FAQ).* Online. Available: http://www.watermarkingworld.org/faq.html (Accessed on 12 August 2005)

Weibel, Stuart and Koch, Traugott. (2000). The Dublin Core Metadata Initiative: Mission, Current Activities, and Future Directions. *D-Lib Magazine*, 6(12). Online. Available: http://www.dlib.org/dlib/december00/weibel/12weibel.html (Accessed on 26 September 2001)

Weibel, Stuart and Miller, Eric. (1997). Image Description on the Internet. A Summary of the CNI/OCLC Image Metadata Workshop, Dublin, Ohio, 24-25 September 1996. *D-Lib Magazine*, January. Online. Available: http://www.dlib.org/dlib/january97/oclc/ 01weibel.html (Accessed on 13 August 2001)

Wert, Carlos and Hernandez, Francisca. (2001). COVAX Project. *Cultivate Interactive*, 3. Online. Available: http://www.cultivate-int.org/issue3/covax/ (Accessed on 1 June 2002)

Whitelaw, Alan and Joy, Gill. (2001). *Summative Evaluation of Phrase 3 of the eLib Initiative: Final Report.* Guildford, Surrey: ESYS. Online. Available: http://www.ukoln.ac.uk/ services/elib/papers/other/summative-phase-3/elib-eval-main.pdf (Accessed on 1 September2002)

Widom, Jennifer. (1999). Data Management for XML. *IEEE Data Engineering Bulletin, 22* (3): 44-52.

Williams, Kevin et al. (2000). *Professional XML Database.* Birmingham, UK: Wrox Press, Chapter 1.

Willis, Katherine. (1995). TULIP at the University of Michigan. *Library Hi Tech,* 13(4): 65-68.

Wittern, Christian. (1995). The IRIZ KanjiBase. *The Electronic Bodhidharma*, (4): 58-62.

Wittern, Christian. (2000). *SMART Project: Methods for Computer-Based Research of Premodern Chinese Texts.* Paper presented in the ACH/ALLC Conference, University of Glasgow, 21-25 July 2000.

WorldLanguage.com. (2004). *Fine Reader 7.0 OCR Pro.* Online. Available: http://www.worldlanguage.com/Vietnamese/Products/104256.htm (Accessed on 16 November 2004)

XML-Signature Syntax and Processing. (2002). Online. Available: http://www.w3.org/TR/ xmldsig-core/ (Accessed on 26 September 2003)

XML-Signature XPath Filter 2.0. (2002). Online. Available: http://www.w3.org/TR/ 2002/REC-xmldsig-filter2-20021108/ (Accessed on 26 September 2003)

Yang, Shu Ching. (2001). An Interpretive and Situated Approach to an Evaluation of Perseus

Digital Libraries. *Journal of the American Society for Information Science and Technology,* 52 (14): 1210-1223.

Zorich, D M. (2003). *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns*. Online. Available: http://www.clir.org/pubs/reports/pub118/contents.html (Accessed on 16 June 2004)

Appendix I

Interview Questionnaire

This appendix lists the questions used in our case study which provides the data used in the research discussions which form part II of this thesis.

Rationale for Digitization

Q. Why are you digitizing material?

Q. What types of collections have been selected for digitization?

Q. What factors influence the content selection and evaluation, from the practical and technical points of view? Does XML technology influence the selection policy?

Q. What demands (such as meeting preservation needs and increasing access) or institutional strategies (such as reducing costs and attracting funding) will be met by digitizing them?

Realizing the Digital Libraries

A. The Digitizing Process

Q. What materials are suitable for digitizing?

Q. Do you use optical character recognition (OCR) to create transcriptions of textual documents?

Q. What materials are suitable for text encoding?

Q. Is the text-encoding done by human labour or machine?

Q. What have you done about quality assurance?

Q. What are the issues regarding hardware, software and resulting file formats?

B. Metadata and Metadata Systems: XML Application

Q. What metadata formats have been implemented for digitization?

Q. How do you manage the consistency of metadata formats?

Q. Does your metadata system use XML technology?

Q. What is the level of detail in the metadata tagging?

Q. Do existing metadata schemas work well? If not, what do you do about this? Do you create your own?

C. Access Management

Q. Are you working on the area of authentication and authorization, and what is the current approach?

D. Delivery System

Q. Do you use an in-house developed system or a vendor's system? Is this an XML-aware system? How does the system handle text and non-textual data?

Q. Do you use the Resource Description Framework (RDF) technology? How is RDF implemented in your system? Does it link to other systems?

Q. What type of database do you use?

Q. How many computing personnel are involved in setting up and maintaining the delivery system?

Maintaining the Digital Libraries

A. Managerial Aspects

a. Staff infrastructure

Q. How has staff structure changed due to digital library development, especially XML-based? What are the implications for staff recruitment, retention and development?

Q. What kind of people do you need to hire to do the work, particularly tasks related to XML technologies?

Q. Do you provide staff training? What skills do they need?

b. Cost centre structure

Q. Where is the money spent? Which areas are the significant parts of the cost?

Q. What is the cost of development and of maintenance?

c. Maintenance and future development

Q. What is the growth rate? Could you please provide statistics?

B. Aspects of Usage

Q. What is the benefit to users of XML?

Q. What feedback do you have from users?

Q. Have you developed an evaluation scheme for digital library services and collections?

Q. What kind of approach to DL evaluation have you developed (for example, evaluating systems, interfaces, or evaluation from the users' point of view - user needs, preferences and user community)? How far have these gone?

C. Organizational Aspects

a. Institutional strategy

Q. What partnerships and collaborations do you have? What is the rationale for these?

Q. What are your models for sustainability? Where did the original funds come from?

b. Future work

Q What are the major contributions of your Digital Library?

Q. What are the future directions?

Appendix II

User Survey Questionnaire

Survey of Usage of Digital Libraries

This survey is investigating user satisfaction on the Perseus Digital Library/University of Michigan Digital Library/Library of Congress National Digital Library (American Memory). This research is part of a PhD thesis for Library and Information studies at University College London, UK. Thank you for your time and your contribution to my research. Please send completed questionnaire to uczcncc@ucl.ac.uk.

Naicheng Chang
UCL

___ **1. Your gender:**

a) male
b) female

___ **2. Your age:**

a) -15 years
b) 16-18 years
c) 19-25 years
d) 26-35 years
e) 36-45 years
f) 46-55 years
g) 56+ years

___ **3. Reason for researching topic:**

A) General interest/lifelong learning

___ **B) School project/academic research (Please choose level)**

a) high school
b) undergraduate
c) master's/doctoral
d) adult/continuation education

___ **C) Classroom instruction**

a) elementary school
b) middle school/junior high
c) high school
d) undergraduate
e) master's/doctoral
f) adult/continuation education

D) Other (please explain)

___ **4. How did you learn about the Digital Library?**

a) Recommended (such as teachers, friends, colleagues etc.)
b) Found on Internet
c) Recommended in printed material, e.g. book, article
d) Other (please state):

___ **5. Have you ever had any training?**

a) Demonstration in a group
b) Personal tuition
c) Through use of HELP features
d) None

___ **6. How often do you use the Digital Library?**

a) Once ever
b) Occasionally
c) Several times a year
d) Several times a month
e) Several times a week
f) At least once a day

___ **7. Are you satisfied with the response time?**

a) yes
b) medium
c) no

___ **8. What features which are not included would improve your experience?**

a) Simple Boolean searching
b) Advanced Boolean Searching
c) Contextual help (help tailored to the page you are at)
d) HELP opening in a new window (HELP messages in a new window would be easy to close and return to the page where you were before)
e) Improved presentation for linkages
f) Online tutorial with screens depicted
g) FAQs
h) Others (Please specify)

___ **9. How do you find the Digital Library search interface?**

a) easy
b) middle
c) hard (Please specify any problems)

10a. In the case of Perseus, have you found Perseus Tools useful in finding information?

Perseus has 16 tools which could be categorized as **A)** General Index, **B)** Image Index, **C)** Map, **D)** Dictionary, and **E)** Virtual Reality. For each please evaluate below:

___ **A) General index (Collection Viewer, English Index, Perseus Table of Contents)**

a) very useful
b) medium
c) no use
d) not used

___ **B) Image index (Art & Archaeology Browser)**

a) very useful
b) medium
c) no use
d) not used

___ **C) Map (London Atlas)**

a) very useful
b) medium
c) no use
d) not used

___ **D) Dictionary (Dictionary Entry Lookup, English to Greek Word Search, English to Latin Word Search, Greek Morphological Analysis, Greek Vocabulary Tool, Greek Words in Context, Latin Morphological Analysis, Latin Vocabulary Tool, Latin Words in Context, Lookup Tool)**

a) very useful
b) medium
c) no use
d) not used

___ **E) Virtual Reality (Virtual Reality Interface)**

a) very useful
b) medium
c) no use
d) not used

___ **10b. In the case of Michigan Digital Library, have you found the search interface useful in finding information?**

a) Search by individual collection
b) Basic search within collection by different fields (full text, title, etc)
c) Basic search within collection by keyword in any field
d) Advanced search (using Boolean, proximity, etc)

10c. In the case of the Library of Congress American Memory, have you found the different kinds of search interface useful in finding information?

___ **A) Search all collections**

a) very useful
b) medium
c) no use
d) not used

___ **B) Browse Collections by Topic**

a) very useful
b) medium
c) no use
d) not used

___ **C) Browse "Collections Containing" (e.g. form of digitized material)**

a) very useful
b) medium
c) no use
d) not used

___ **D) Browse Collections by Time Period**

a) very useful
b) medium
c) no use
d) not used

___ **F) Browse Collections by Place**

a) very useful
b) medium
c) no use
d) not used

___ **11. Does the Digital Library provide enough help in getting started or in finding what you want from the data?**

a) yes
b) medium
c) no

___ **12. What are the major problems while searching the Digital Library?**

a) too many links
b) get lost easily
c) retrieve wrong material
d) Other (please state)

___ **13. Do you find it easier to use for some different fields rather than others (e.g. history, religion, literature, science etc)?**

a) yes (please mention which field)
b) no
c) don't know

14. Are there any additional comments you would like to make?

(please state)